Panel Data Sample Selection Model: an Application to Employee Choice of Health Plan Type and Medical Cost Estimation

Jeonghoon Ahn*
Dept. of Pharmaceutical Economics & Policy
University of Southern California

Abstract

This paper utilizes a nonparametric panel data sample selection model to correct selection bias in the analysis of longitudinal medical claims data. Selection bias in the health economics data is a common problem and many health economists have used Heckman type selection models in cross-sectional analyses. Since longitudinal data structure is common in health economics data, especially medical claims data, the correction of selection bias in the longitudinal sense is especially valuable for health economics related researches. The complicated modeling and extensive computer programming needs, however, resulted to only a few health economics researches in this direction. This paper suggests a relatively simple estimation framework to correct sample selection bias in longitudinal data. An example of health care utilization of PPO type plan holders in an employee pool is also provided as follows: in the first step, a random effect panel data probit model was used to estimate each employee's choice between HMO type plans and PPO type plans; in the second step, a nonparametric fixed effect panel data selection model, using the estimates from the first step, was used to estimate the medical expenditures of PPO plan holders (similar to Kyriazidou, Econometrica 1997). Since the second step estimation can be expressed as a weighted least squares regression, this framework is simple to use, but among others, this nonparametric framework is robust from any parametric misspecification and free from a controversial health econometric problem called retransformation in two part model (Manning, Journal of Health Economics 1998; Mullahy, Journal of Health Economics 1998; Ai and Norton, Journal of Health Economics 2000). There are some interesting results from this example, but among others, the selection bias influenced significantly on the Age effect of medical expenditures. Since there were more young employees in the HMO plan holders, the Age effect of PPO plan holders was almost doubled after considering for selection bias.

JEL classification: I11; C23; C25

Keywords: Medical Expenditure Estimation; Panel Data; Probit Regression; Risk Adjustment; Sample Selection Model

* Jeonghoon Ahn, Assistant Professor, Department of Pharmaceutical Economics and Policy, University of Southern California, 1540 E. Alcazar St. CHP-140, Los Angeles, CA 90089-9004, U.S.A. Tel.: +1-323-442-1461; fax: +1-323-442-1460. *E-mail address:* jeonghoo@usc.edu.

1 Introduction

Panel (longitudinal) data is a stack of cross-sectional data observed over time for the same cross-sectional units, such as individual identifiers. Since panel data enables us to follow an individual's history over the time periods, it is a natural extension of cross-sectional data which is only observed at a specific time period. For a simple rationale for panel data use, consider an example[1] of studying health plan type choice between HMO and non-HMO; a cross-sectional data (say year 1996) showing a half of the population selected HMO and the other half did not. This finding can be interpreted as (A) each individual has an identical and independent probability of 50% to choose HMO type plan, or (B) a half of the population is born to choose a HMO plan and the other half is born not to choose a HMO plan. The case (A) implies that frequent switching behaviors are expected whereas the case (B) implies that no switching behaviors are expected at all. Obviously, both of the interpretations are two extreme possibilities (probably the reality is somewhere in between these two cases), however, we cannot empirically deny these possibilities unless we have another cross-sectional data (say year 1997) of the same population, that is a panel data structure.

In addition, the methodology used for panel data has some benefits for two important problems of cross-sectional data analysis; unobserved heterogeneity and omitted variable bias. Since a typical cross-sectional data analysis is built on the homogeneity of the given sample, unobserved heterogeneity is always a potential critique for most cross-sectional analyses. On the other hand, panel data model can control for unobserved heterogeneity by parametrizing it as either fixed effect or random effect. Furthermore, we can test the validity of heterogeneity by comparing a panel data model and a corresponding pooled cross-sectional data model. For random effect modeling, a pooled cross-sectional model assumes no correlation in the error

[1] See Ben-Polath (1973) for the original example of female labor force participation.

term as in the classical linear regression theory, whereas a random effect panel data model allows correlated error terms for the observations from the same individual. Therefore, the pooled cross-sectional model is a nested specification of the random effect panel data model and a test can be easily set up by using either error sum of squares (F-test) or log likelihoods (χ^2 test) depending on the estimation scheme. Similar tests are also available for fixed effect modeling.[2]

Controlling for unobserved heterogeneity helps to achieve more accurate prediction. As an extension to the previous example of health plan type choice, we can consider a panel data model with two years 1996 and 1997. This model allows Jane Doe's choice at 1996 and her choice at 1997 to be correlated and so to be John Doe's choice at 1996 and his choice at 1997.[3] Hence, if we admit the past choices are useful information to predict the current choice, panel data model is at least intuitively appealing methodology (we will have a perfect forecast for the 1997 choices if the case (B) is true). On the other hand, the panel data model is no better than the pooled cross-sectional model if the case (A) is true.

Another benefit of panel data analysis methodology is solving omitted variable bias, unless the omitted variable is time varying. Once we control for the unobserved heterogeneity by either fixed effect or random effect modeling, we cannot distinguish anything time invariant from fixed effect or random effect parameter. Therefore, we do not have to worry about any time invariant omitted variable[4]. On the other hand, this benefit is also a weakness of panel data model since we need additional stratifications of the analysis whenever we are interested in the effect of observable time invariant variable such as gender, race, etc. For example,

[2] For example, Greene (1997) p.617.

[3] If we have more than 3 time series observations ($T \geq 3$), we can use panel data model with a time series error structure to refine these correlations to be a causal relationship, i.e. only the past affects the future not vice versa.

[4] This can be very useful in health economics research; for example, the general health of individual, say healthy (low risk profile) vs. non-healthy (high risk profile), cannot be observed but need to be considered in the model then panel data model can be a good solution.

if the health plan type choice difference between female employees and male employees is a question of interest, we need two separate panel data models by each gender to compare them[5].

Selection bias is an inevitable problem in many empirical researches, especially for the ones using a retrospective data. Since the retrospective data is not randomized for the research objective, any data-driven conclusion cannot be generalizable beyond the data set. Therefore, the conclusions which might be closely related to the data selection process, is subject to selection bias. Even though there are difficulties in practice, the importance of selection bias in panel (longitudinal) data is no less than in the case of cross-sectional data. Especially, many health economics data set has a panel data structure and the observability is related to the individual choice, for instance, a typical health insurance claims data.

Selection bias problem was extensively considered in the health plan type choice between Health Maintenance Organization (HMO) and Preferred Provider Organization (PPO), or between HMO and Fee For Service (FFS). For instance, Eggers (1980), Dowd, *et al*. (1996), and Riley, *et al*. (1996) were the selection bias studies on Medicare enrollees. Also Hellinger (1995) had a nice review article on selection bias studies related to HMO vs. PPO choice in general. Most of these studies showed that HMO plan enrollees are healthier than PPO or FFS plan enrollees, which implies that any projections solely based on one type of plan is subject to a selection bias.

Selection bias modeling is also closely related to the concepts of "risk adjustment[6]" and "adverse selection." Risk adjustment is nothing but capturing different health status (risk) of individuals to explain

[5] In a cross-sectional analysis, we can also test the coefficient differences between two gender group by Chow test but it is impossible in panel data analysis unless jointly testing coefficient differences and validity of heterogeneity.

[6] Risk adjustment captures selection bias by new variables such as Adjusted Average Per Capita Cost (AAPCC), Principal Inpatient Patient-Diagnostic Cost Group (PIP-DCG), Diagnostic Cost Group (DCG), Hierarchical Condition Categories (HCC), etc. Therefore, these methodologies can be viewed as correcting an omitted variable (true health status) bias by a set of instrumental variables in econometrics language.

outcomes such as medical expenditures, mortality rate, and so on. Therefore, these studies seek for a remedy of more general type of bias[7] in terms of missing information. Hence, various risk factors are identified and many different risk measures are developed in this line. Iezzoni (1997) and Van de Ven and Ellis (2000) summarized various studies related to this issue. Adverse selection or hidden information is more familiar word for the economists and there are also many researches in this aspect, for example, van de Ven and van Vliet (1995), Neudeck and Podczeck (1996), Ettner (1997), Altman, Cutler and Zeckhauser (1998), and many others.

Another interesting application of selection bias modeling is on medical expenditure estimation. Since a medical expenditures data set is typically skewed and the coefficient can be easily interpreted as an elasticity, a log transformation of expenditure variable is widely used[8]. However, we cannot take logarithm on zeros. Therefore, we need a model to separate zero expenditure observations and positive expenditure observations, such as two part model (Cragg, 1971) or sample selection[9] model (Heckman, 1976, 1979). The two part model estimates a selection equation and a main equation independently whereas the sample selection model considers both equations jointly. If the proportion of zeros is small, one part model (Duan, 1983) or a model without a log transformation can be used. Also there are important estimation issues arising if the proportion of zeros is small, for instance, the choice of probit model in the first stage of sample selection model and two part model becomes significant.

---

[7] Risk adjustment is dealing with mostly individual risk or closely stratified group's risk. However, it is equivalent to selection model if we consider different risk of two groups separated by a selection equation.

[8] In addition, there could be an efficiency gain from reducing noises by using a lognormal specification. Duan *et al*. (1983) said the HIS data achieves this gain (the relative efficiency of log transformed model over raw model is greater than one).

[9] Selection bias can be modeled in many different ways, but one way to classify selection bias models is separating by sample selection models and self selection models. According to Maddala (1985a), a sample selection model employ a selection equation written in a reduced form while a self selection model employ one in a structural form.

There were serious debates on the choice between two part model and sample selection model (Duan *et al*., 1983, Hay and Olsen, 1984, Duan *et al*., 1984, Maddala, 1985a, Duan *et al*., 1985, Maddala, 1985b). Since the two part model, which has a long history, was adapted as the empirical model for the RAND Health Insurance Experiment (HIE), the same RAND researchers were actively participated for the advocacy of two part model (Jones, 2000). However, many econometricians believe selection bias is an important problem to be considered. The two part model seemed to have an edge over the sample selection model, since Manning *et al*. (1987) and Hay *et al*. (1987) found the better performance of the two part model through a Monte Carlo simulation. However, Leung and Yu (1996) showed these simulation results does not hold for a simulation design with enough variations in independent variables. Hence, they argued that the specific simulation design, which created the collinearity between the inverse Mill's ratio and the variables in the main regression, generated favorable result for the two part model against the sample selection model in Manning *et al*. They also mentioned that it may also be the case in Hay *et al*. Hence, it is more meaningful to highlight distinctive advantages of each model than to argue one model is better than the other.

For a cross-sectional data, Duan *et al*. (1984, 1985) made a good distinction between a two part modeling and sample selection modeling. The former used for the Health Insurance Experiment (HIE) is better for a conditional question of their interest, the average medical expenditures of the people who have spent nonzero amount (subject to selection bias if we want to interpret the result for everybody in the sample). On the contrary, the econometricians are more interested in an unconditional question, the average medical expenditures of everybody in the sample including people with zero expenditure (selection bias is corrected). For panel data, the conditional and unconditional question distinction of Duan *et al*. is not very useful, since

the non-zero expenditure population is varying over time. For instance, panel data two part modeling has to treat this individual differently for each year if an individual had zero expenditure in one year and some positive expenditure in the next year. Therefore, two part model is not necessarily even preferred to one part model in panel data. Jones (2000) also made a good distinction between a two part model and a sample selection model. He recommended two part model for "genuine zeros" and sample selection model for non-observable responses. The term "genuine zeros" implies that zero observations in a dependent variable is not missing observations, whereas non-observable responses refer to missing observations. Non-observability can be resulted from many different reasons, for instance, deductible truncates observable claims data to be above deductible level. In our data, non-observability of medical expenditures was from each individual's choice of the health plan type between HMO and PPO. Therefore, this data is more suitable for a sample selection model according to Jones (2000).

In addition to the comparison with two part model, there are some additional differences between cross-sectional sample selection model and the panel data sample selection model in this paper. Especially, the nonparametric panel data selection model in this paper is free from a critique by Duan et al. (1983, 1984), Heckman's sample selection model is dependent on bivariate normality assumption. Since we are considering a nonparametric estimation of the main equation coefficients, we do not need a distributional assumption between the selection equation and the main equation nor a parametric specification of the selection equation.[10] This advantage of nonparametric estimation is especially important for applied researches, since a misspecification problem can always affect the validity of conclusion but a nonparametric specification is less vulnerable to any misspecification problem than a parametric specification. However, a large data set

[10] Instead, we need a mild conditional exchangeability condition (See footnote under equation (6) in Model section for more details).

is required for nonparametric estimations due to a slow convergence rate in asymptotic normality.

Another important issue related to a panel data selection model for medical expenditure is regarding heteroskedasticity and log transformation. The log-transformed dependent variables have to be retransformed for the easier interpretations and policy conclusions. However, the retransformation formula (for example, Duan (1983)) includes the variance parameter which cannot be consistently estimated by an ordinary model under heteroskedasticity. Manning (1998) showed various cases with the bias resulted from heteroskedasticity and there were also some suggestions for alternative modeling (Mullahy, 1998, Ai and Norton, 2000). This is especially important for panel data sample selection models and even for the cross-sectional sample selection models using LIML estimation (Manning, 1998), since these models introduce heteroskedasticity in their selection bias correction procedures. To summarize, the costs of using log transformation in panel data selection models need additional steps to recover correct retransformed values and careful considerations for interpretation (since zero expenditure population is varying).

The two important estimation issues for empirical health economics, panel data and selection bias, are traditionally treated separately (not only in health economics). Since it is mathematically complex to combine these two issues together, a large burden of computer programming and a set of strong distributional assumptions are need for the combination. The model presented in this paper can be estimated with a common statistical software such as STATA or LIMDEP.[11] Also the statistical assumptions needed for the model in this paper is relatively weaker than the other methods.

This paper is organized as follows, Section 2 explains the data for our analysis, Section 3 introduces a panel data model of health plan type choice and a nonparametric panel data sample selection model of

[11] Unfortunately, there is no specific command to perform the whole model presented here, but a relatively simple programming can achieve the correct estimation results. Every estimation results in this paper is generated by STATA 7.0.

medical expenditure, Section 4 summarizes estimation results, and Section 5 concludes with some remarks.

2 Data

This paper used a data set provided by the University of Southern California (USC) benefit office. This data set includes two subsets, the employee data and the claims data, they can be connected through encoded ID number. The claims data is only observable for the people who chose PPO plan. Therefore, any conclusion based on the claims data is only valid for PPO plan holders at the specific time and any extension of this conclusion to all the employees is subject to selection bias. The time range of data set is two years, 1996 and 1997. There are 8,543 employees for the year 1996, 8,596 employees for the year 1997, and 12,615 employees for the two years (some of them have only one year data). These numbers include people who chose only supplemental coverage or cash compensation for their already existing outside health insurance coverage (may be covered by their spouse's health plan). Since we cannot observe these outside health plan characteristics, these people are not included in the analysis. After excluding these outside plan holders and coding error data, we get 7,743 employee data and 7,762 data for year 1996 and 1997, respectively.[12] Combining these two data yields 6,644 employees in a two year balanced panel data and 8,861 employees in the two year unbalanced panel data (balanced panel data plus 1,099 employees for the single year 1996 only and 1,118 employees for the single year 1997).

Available health plans were three HMO plans and two PPO plans, all five plans offered in both years. The two PPO plans are offered by the university network, which includes the services by USC faculty physicians

[12] There were seven coding errors for the age variable, two coding errors for the experience variable and one data for 1998 which might be a coding error of the date variable.

at the USC hospitals[13], while the three HMO plans are all outside organizations: Kaiser Permanente (KP), CaliforniaCare (CC)[14] and Pacificare (PC)[15]. The two university network plans are PPO type plans but the coverage and the preferred rates are different: one is a basic coverage plan (NET 1) and the other is a more extensive coverage plan (NET 2).

There were two major changes in the design of health plans offered by the university in 1997. Both changes were applicable to university network plan holders only. The first change was made on the deductible of university network plans (PPO): there was no deductible for the direct services from the university in 1996 but $100 per person and $300 per family annual deductible was introduced from 1997. Also the annual deductible of services rendered by all the other preferred provider groups in the university network plan, was increased from $100 to $150 per person and from $300 to $450 per family. The second change was the increase in premiums of the basic coverage university network plan (NET 1) while all the other plan premiums stayed the same. The limitation of this data set is that we cannot distinguish any effects of the deductible change and the premium change since we are using a data for only two years (two different benefit designs).

In Table 1, we can see that there is not much difference in means between whole data and balanced panel data. Only one notable difference is the mean of the D_{HMO} variable in the 1997 whole sample and the mean of the 1997 balanced sample, which indicates that the dominant number of newly hired employees of 1997 data chose HMO as their first health plan to start at USC. This makes sense since if they do not have enough information about each health plan then why not start with a less expensive plan? One important variable we

[13] USC hospitals have more than 2000 beds and it is one of the biggest teaching hospital.

[14] This is a Blue Cross/ Blue Shield(BCBS) descendent.

[15] The business of this plan is somehow connected to CaliforniaCare so that the provider network is identical for the USC employees.

do not have is the household income or household wealth. Since we do not have even employee salary, we leave it for the panel data model to deal with as an omitted variable. If this variable did not vary much from 1996 to 1997, our result is free from the omitted variable bias. Also some results from the previous literature show the effect of income on the choice of health plan type is small (Barringer and Mitchell, 1994). Table 2 shows the basic structure of health plan choices for the unbalanced panel data. If we consider the balanced panel data only (for 6,644 employees), only 3.3 percent of employees switched their plans to another plan and 2.4 percent of employees switched the type of the plan (i.e. HMO vs. PPO). From Table 3, we can see the coverage structure is very stable for those who use this health benefit,[16] 94% of employees in balanced panel data (the first nine cells from the upper left corner) did not change coverage type and 80% of the all employees did not changed their coverage choice from 1996 to 1997. The biggest two changes in balanced panel data are from single coverage (1, SINGLE) to employee plus one dependent coverage (2, PLUS ONE) and employee plus one dependent coverage category to family coverage (3, FAMILY).

3 Model

Similar to the Heckman's cross-sectional sample selection model (Heckman, 1976, 1979), this model also has two steps. The first step is to estimate a health plan type choice (selection equation) and the second step is to estimate medical expenditures (main equation). An example diagram showing the idea is in Figure 1. There might be some variables affecting both the choice of health plan type choice and the medical expenditure, but they are not required to be included. We included only the statistically significant variables in the estimation results. Another application of the model is presented in Figure 2. As the health

[16] People who are not associated with outside option, i.e. balanced panel data.

plan type choice above, a prescription fill up decision can truncate a sample of drug expenditures. In this case, a selection equation of prescription fill up decision and a main equation of drug expenditure can be modeled similarly. Grootendorst (1997) applied a panel data tobit model on the drug expenditures which is truncated by deductible limit not by a selection equation.

In the first step, a reduced form econometric model for multi-period health plan type choice between HMO vs. PPO can be written as follows,[17]

$$y_{it}^{*} = \beta x_{it} + \varepsilon_{it}, \quad i = 1, ..., N \; ; t = 1, ..., T_i \tag{1}$$

$$\begin{aligned} y_{it} &= 1 \text{ if } y_{it}^{*} > 0 \\ &= 0 \text{ otherwise,} \end{aligned} \tag{2}$$

where y_{it}^{*} is the unobserved propensity to join a HMO type plan by individual i at time t, and y_{it}.= D_{HMO} denotes the choice of health plan type whether HMO (D_{HMO}=1) or PPO (D_{HMO}=0), x_{it} is a k-dimensional vector of observable variables including demographics and health plan characteristics. ε_{it} is an error term. T_i = 1 for all i implies usual cross-sectional Limited Dependent Variable (LDV) model, which is commonly used in health economics literature. For example, Ellis (1985) used a logit model to analyze employee plan choice, Hornbrook *et al.* (1989) examined selectivity and selection bias using a probit version; Buchmueller (1995) used a probit model for comparing the effects of employer provided health insurance types on health status. Also Buchmueller and Feldstein (1997), employed a probit model to study employees' behavior of switching health plans (a dichotomous variable of switch as the dependent variable). Other health economics applications of cross-sectional probit model can be also found in log-transformed med-

[17] There are many structural form models which can result in this reduced form model. For example, a choice model between price vs. quality can be well fit into the situation (HMO vs. PPO).

ical expenditure estimation with a good proportion of zero expenditures (two part model and Heckman's sample selection model both use probit regression in the first step).

For the panel data estimation ($T_i > 1$ for some i), a random effect probit model is considered by the following error structure[18] on ε_{it},

$$\begin{aligned} \varepsilon_{it} &= u_i + v_{it} \\ u_i &\sim N(0, \sigma_u^2), \quad v_{it} \sim iid\, N(0,1), \quad u_i \perp v_{it} \end{aligned} \tag{3}$$

where u_i is an unobserved heterogeneity among individuals (which is time invariant), v_{it} denotes an underlying innovation, and $\perp$ indicates independent relationship. This model is more general than pooled cross-sectional model since the error terms for each individual are correlated through the common u_i (every error terms are independent in pooled cross-sectional model). In econometric sense, allowing intertemporal correlation for each individual (panel data model) is better than including the past choice (y_{it-1}) in the equation (a modification to cross-sectional model), since the latter introduces endogeneity bias.

An alternative panel data binary choice model can be found in two directions; fixed effect modeling or logit specification. However, Chamberlain (1984) and Hsiao (1986) explained that the fixed effect probit models do not provide a consistent estimator of β since there is an incidental parameter u_i. This notorious incidental parameter problem resulted from the fact that the number of parameter to estimate increases faster or at the same rate as the sample size N increases.[19] Logit specification is an attractive choice since

[18] In this paper, identifying restrictions are already imposed on the variance of innovations ($\sigma_v^2 = 1$). Alternative identifying assumption can be made as $\sigma_u^2 + \sigma_v^2 = 1$.

[19] Generally, increasing the sample size N to a large enough number yields an efficiency gain, but the incidental parameter problem is an exception.

it provides both fixed effect[20] and random effect models, however, the choice between logit and probit, is not simple in panel data analysis.[21] Unlike the cross-sectional data case (a univariate distribution), normal distribution (probit) and logistic distribution (logit) is not that similar in panel data (a multivariate distribution). Therefore, the relative performance of each specification is the only measure for the specification choice. Since the probit specification showed a better prediction of the actual choices in our data, we did not use a logit specification.

According to Mundlak (1978), random effect model can be viewed as an inference with respect to population whereas fixed effect model can be viewed as an inference conditioning on the given sample as a draw from the population. This guided our choice of panel data modeling; random effect for the health plan type choice and the fixed effect for the medical expenditure model, we believe we have enough variables to model the health plan type choice (which is also a relatively simple problem) for the whole population but we are less certain for medical expenditure since there are many factors affecting the estimation procedure such as distributional assumption, heteroskedasticity, and so on.

To estimate above equation, we follow Butler and Moffit (1982). This method utilizes Gauss-Hermite

[20] The Conditional Maximum Likelihood (CMLE) of Chamberlain (1980) can be used for the fixed effect logit model, but it is inefficient since it does not reflect any consistent choices (i.e. people who did not switch health plan type), which is more than 96%.in our balance panel data.

[21] Only the tail probability is slightly different in cross-sectional case, so the choice of probit or logit does not yields significantly different result. Amemiya (1981) showed a conversion formula between the coefficients from each specification.

quadrature[22] to evaluate inner integrals in our likelihood function.[23]

$$\begin{aligned} L &= \Pi_{i=1}^{N} P[y_{i1} = b_{i1},\, y_{i2} = b_{i2}, ...,\, y_{iT_i} = b_{iT_i}) \\ &\approx \Pi_{i=1}^{N} \frac{1}{\sqrt{\pi}} \sum_{k=1}^{K} w_k g(a_k) : \text{Gaussian-Hermite Integration} \end{aligned} \tag{4}$$

where $b_{it} = 0$ or 1 only, w_k is a quadrature weight, a_k is called a quadrature abscissa, $g(r_i) = \Pi_{t=1}^{T_i} \Phi[(2d_{it} - 1)\ (\beta x_{it} + \sigma_u \sqrt{2} r_i)]$, and Φ is the standard normal cumulative distribution function. Once w_k's and a_k's are provided, this likelihood function can be easily maximized with respect to β.[24] Generally, we need higher K to achieve good approximation as T_i grows higher and $\rho = \sigma_u^2/(1 + \sigma_u^2)$ gets larger. Butler and Moffit(1982) method has a simple implication on testing the existence of random effect. Since there exists an equicorrelation[25] parameter due to the random effect coefficient (σ_u^2 in equation (3)), we can test the significance of this equicorrelation parameter to judge whether there exists a valid random effect. The estimated results are presented in Table 4 and Table 5.

A panel data analysis model considering both unobserved heterogeneity and selection bias has to overcome severe[26] nonlinearity and nuisance parameter problem in the model. In the panel data sample selection models, maximum likelihood estimation needs more statistical distributional assumptions since unobserved heterogeneity parameter (such as u_i) also appears in the distribution of underlying error term (ε_{it}). For

---

[22] Gaussian-Hermite Integration formula:

$\int_{-\infty}^{\infty} e^{-r^2} g(r) dr \cong \sum_{k=1}^{K} w_k g(a_k)$

[23] See Ahn (2000) or Greene (1997) for more details.

[24] The tabulated values of w_k and a_k can be found in Table 25.10 of Abramovitz and Stegun (1972).

[25] The random effect panel data model specified as (3) has a multi-period defect called equicorrelation. Since there is no additional parameter introduced to model the decay of correlation as time periods gets further, the intertemporal correlation is same for any two time periods.

[26] Severe in a sense that nonlinearity from selection bias can not be differenced out in a similar way to cross-section data case, i.e. time-varying nonlinear component.

example, a random effect sample selection models need to specify a joint distribution of four arguments: two underlying error terms from a selection equation and a main equation, two random effect (unobserved heterogeneity) parameters (Hsiao, 2001). There have been several solutions suggested for this problem, however, they are either dependent on strong distributional assumptions or parametric specification of the selection equation, which implies they are vulnerable to misspecification problem. Baltagi (1995) has a nice summary on these studies. Kyriazidou (1997) proposed to use nonparametric kernel weight for correcting selection bias, so that no bivariate normality assumption between the selection equation and the main equation, is required. Since this estimation methodology does not require a parametric specification of selection process (typically written as a probit regression equation), it is especially useful for applied researches, where a misspecification is always an issue.

In the second step, we applied aforementioned Kyriazidou's panel data selection model in our medical expenditure estimation as follows,[27]

$$\begin{aligned} E_{it} &= d_{it}E_{it}^{*} \\ &= d_{it}(w_{it}^{*}\gamma + \delta_i^{*} + \xi_{it}^{*}) \\ &= w_{it}\gamma + \delta_i + \xi_{it},\ i = 1...N,\ t = 1, 2 \qquad (5) \\ d_{it} &= I\{x_{it}\beta + u_i - \upsilon_{it} \geq 0\} \qquad (6) \end{aligned}$$

where E_{it} is the medical expenditure of individual i at time t. d_{it} is the health plan type choice variable (=1-y_{it}). E_{it}^{*} is a latent variable only observed for $d_{it} = 1$. w_{it}^{*} is a vector of explanatory variables including health status and age (w_{it}^{*} and x_{it} may have common variables). δ_i^{*} and u_i are unobserved heterogeneity

[27] $t = 1$ is for 1996 data and $t = 2$ is for 1997 data. It can be generalized for $t > 2$.

coefficients but we assume δ_i^* to be a fixed effect parameter while we keep u_i as a random effect parameter.[28] Now We can rewrite (5) as

$$E_{it} = w_{it}\gamma + \delta_i + \lambda_{it} + \omega_{it} \tag{7}$$

where λ_{it} denotes sample selection parameter. $\omega_{it} = \xi_{it} - \lambda_{it}$ and it satisfies $E\ (\omega_{it}|\, d_{i1} = 1,\, d_{i2} = 1,\, x_{i1},\, x_{i2},\, w_{i1}^*,\, w_{i2}^*,\, u_i,\, \delta_i^*) = 0$. If this was a cross-section version, Heckman(1976) can be applied and we substitute λ by the inverse Mill's ratio. To estimate γ consistently, Kyriazidou(1997) proposed the following estimator[29];

$$\hat{\gamma} = [\sum_{i=1}^{N} \hat{\psi}_{iN}\, \Delta w_i' \Delta w_i \phi_i]^{-1} [\sum_{i=1}^{N} \hat{\psi}_{iN}\, \Delta w_i' \Delta E_i \phi_i] \tag{8}$$

where Δ is the difference operator ($\Delta w_i = w_{i2} - w_{i1}$), ϕ_i is a trimming dummy variable defined as $\phi_i = d_{i1} d_{i2}$ so $\phi_i = 1$ implies that individual i chose PPO type plans for both years. $\hat{\psi}_{iN}$ is a weight estimated nonparametrically as it declines to zero as the difference $\Delta\, x_i\hat{\beta}$ increases. More specifically,

$$\hat{\psi}_{iN} = \frac{1}{h_N} K\,(\frac{\Delta\, x_i\hat{\beta}}{h_N}) \tag{9}$$

where K is a univariate kernel density function[30] and h_N is a bandwidth parameter which satisfies $lim_{N\to\infty}$ $h_N = 0$. The intuition of this estimator is "differencing out nuisance parameter nonparametrically." If an observation i satisfies $x_{i1}\beta = x_{i2}\beta$ and ϕ_i = 1, we can easily difference out the sample selection parameter

[28] This is different from Kyriazidou (1997). Her original model assumes fixed effect for both individual specific coefficients. Consequently, her conditional exchangeability condition should be modified as the distribution of (ξ_{i1}^*,ξ_{i2}^*,$\upsilon_{i1}|\, u_{i1}$,$\upsilon_{i2}|\, u_{i2}$) is identical for same vector of (x_{i1},x_{i2},w_{i1}^*,w_{i2}^*,u_i,δ_i^*).

[29] This estimator can be also used for a tobit specification, but x and w should not have any variable in common for a tobit specification.

[30] The precise form should be $K\,(\frac{-\Delta\, x_i\hat{\beta}}{h_N})$, since $\hat{\beta}$ is the coefficient vector obtained from HMO choice probit regression. However, $K(\bullet)$ is a symmetric function, $K(\Delta\, x_i\hat{\beta}) = K(-\Delta\, x_i\hat{\beta})$.

λ_{it}. Therefore, we use a kernel weights to penalize the observations far away from $x_{i1}\beta = x_{i2}\beta$, i.e. the highest weights for $\Delta x_{i1}\beta = 0$ and the weights declines to zero as $\Delta x_{i1}\beta$ increases. This estimator is shown to achieve consistency and asymptotic normality. Since we need the consistency of $\hat{\beta}$ to build the consistency of $\hat{\gamma}$, it is important to check the correlation between x and u in our random effect setup.[31]

One convenient feature of this model is that it can be easily estimated by a weighted least square regression[32] with weights being equal to $\sqrt{|\hat{\psi}_{iN}|}$ and using only the people who enrolled in PPO plans for both years (3260 employees). However, we have to use the White heteroskedasticity consistent standard errors in this estimation since we introduced a heteroskedasticity in selection bias correction procedure. The important advantage of this model is that it does not require a parametric specification of selection process (6); instead it only requires a consistent estimator for the selection process. This consistent estimator can be obtained by other methods such as the nonparametric methods of Manski (1987) or Horowitz (1992).

Also there are important remarks on this estimator. Even though the Kyriazidou's estimator achieves the asymptotic normality convergence arbitrarily close to $\sqrt{N}$ rate, nonparametric estimators have slower convergence than parametric estimators, i.e. nonparametric estimators need a larger data set to achieve same rate of convergence.[33] Closely related to this, there is a disadvantage of using the estimator suggested in this paper; the desired $\hat{\gamma}$ is asymptotically biased. We can choose h_N such that an asymptotically unbiased estimator is obtained. However, the rate of convergence to normality is slower than asymptotically

[31] u_i's can be easily calculated from a panel regression of residuals from a Maximum Likelihood Estimation of (4) on i. In our data, all the correlations between u and each independent variable are lower than 0.4.

[32] For $T = 2$ case, it reduces to a cross-sectional regression (since the estimator only depends on the difference of two periods). For a general $T > 2$, $(T-1)$ dimensinal panel data weighted regression should be used.

[33] For example, a parametric estimator with sample size 100 has the convergence rate to normal, $\sqrt{100} = 10$. To achieve the same rate, for example, Horowitz (1992) nonparametric estimator, which has $\sqrt[5]{N^2}$ convergence rate, needs $\sqrt{10^5} \approx 316$ observations.

biased estimators.[34] Therefore, a bandwidth yielding the fastest rate of convergence to normality is used for estimation purpose and the asymptotic bias is corrected later. To correct the asymptotic bias, Kyriazidou suggested a "plug-in" method similar to Bierens (1987).

$$\widetilde{\gamma} = \frac{\widehat{\gamma} - N^{-(1-\delta)(r+1)/(2(r+1)+1)}\widehat{\gamma}_{\delta}}{1 - N^{-(1-\delta)(r+1)/(2(r+1)+1)}} \tag{10}$$

where $\widehat{\gamma}$ is from (8) with $h_N = hN^{-1/(2(r+1)+1)}$, $\widehat{\gamma}_{\delta}$ is the same estimator with window width $h_{N,\delta} = hN^{-\delta/(2(r+1)+1)}$ instead of h_N, $\delta \in (0,1)$.[35] Note that δ close to 1 implies $\widehat{\gamma}_{\delta}$ is close to the original estimator $\widehat{\gamma}$. Another unsolved problem in the nonparametric estimator is the efficiency. As we actually used the observations chose PPO for the two years in our estimation (3260 out of 8861), efficiency loss is an avoidable problem for panel data selection models.

As a summary, the second step estimation can be done in the following order:

Step 1. Estimate a consistent estimator $\hat{\beta}$

Step 2-1. Run a weighted least square regression to achieve $\hat{\gamma}$

Step 2-2. Run an auxiliary WLS with δ (0.1 suggested) to get $\widehat{\gamma}_{\delta}$

Step 2-3. Using (10) to correct asymptotic bias and report $\tilde{\gamma}$ along with

White heteroskedasticity consistent standard error from Step 2-1

[34] Compare to the case of variance estimator in Maximum Likelihood Estimator. This estimator is biased but more efficient than Ordinary Least Squares Estimator.

[35] In our estimation, we used a sixth order bias reducing kernel $K_6(\cdot)$, $h = 1$, $\delta = 0.1$ from her original paper.

4 Estimation Results

We used the unbalanced panel data of 8,861 employees in the analysis. As mentioned in the previous section, our model is composed of two different steps: the first step of health plan type choice and the second step of medical expenditure estimation using a sample selection model. This methodology enables us to draw a conclusion for an average employee whether she or he choose HMO type plan or PPO type plan.

In the analysis, the premium variable is normalized by the cheapest premium in the same coverage category so that the amount reflects opportunity cost to switch to the cheapest plan (RPrem: relative premium).[36] For the PPO plan holders, the RPrem can be interpreted as willingness to stay in the current PPO plan.[37] In the estimation, all the plan characteristics of each plan are represented by premium of each health plan since they are perfectly correlated with each other for given number of dependent coverage. This was pointed out as a defect of using single firm analysis in Feldman *et al.* (1989). So we cannot consider separately price elasticity and cross price elasticity typically in this kind of single firm data. This is why we need to generate a relative premium variable which is a combination of both own premium and the other type plan premium.

We used the number of quadrature approximation point, $K = 30$ to estimate the random effect panel data probit model (4). The estimation results are shown in Table 4. The effect of relative premium is negative, as expected, and the square term is positive but significantly small. These can be summarized as there is a negative nonlinear effect of relative premium (in a reasonable range) on the propensity to choose HMO type plan. The experience term shows a small preference for HMO plans among employees with longer

[36] An alternative can be normalizing by the average premium of HMO plans for the same coverage category, but this may generate unwanted negative values for the cheaper HMO plan holders.

[37] Ahn (2000) used this property to show a simulated probability with respect to the change in willingness to pay amount.

experience at USC. From the negative coefficient of AGE variable, we can infer that the aged employees prefer PPO type plans. The effect of coverage (a proxy for the number of insured) shows strong preference to HMO type. To summarize, young employees with a large number of dependents prefer HMO type plans, which is similar to Ellis (1985) and many others from health economics literature. Also this result fits well to a price vs. quality comparison in health plan type choice. A young employee with many dependents prefers a HMO type plan (price advantage) while a old employee without any dependent prefers a PPO type plan (quality advantage). For the validity check of random effect parameter, a Likelihood Ratio (LR) test on the significance of ρ was performed. The comparison between the maximized likelihood of pooled cross-sectional probit model and the one from proposed random effect panel data model are used to build the LR test (two times the positive difference between the two likelihoods) on ρ. If ρ is significantly different from zero, this supports the existence of random effect. We found that random effect parameter was highly significant. Ahn (2000) showed an asymmetric behavior between HMO and PPO plan holders by simulated probabilities of switch by the increase of relative premium and deductible.[38] However, this was based on only two years data without any significant change, the probabilities for a large premium and deductible increase is questionable.

Table 5 shows the prediction success table of this model. It seems that the panel data probit model predicts almost perfectly whether the choice of health plan type is PPO or HMO. However, this is not that surprising since only 162 people actually switched their health plan types (and we missed about 50). From the example in the introduction section, we knew this data set is close to Case (B) and the correlation between the current choice and the past choice plays a great role in the prediction.

[38] There is a limitation that we cannot distinguish an increase in premium and an increase in deductible, since both of them are changed at once and we have only two years data.

Two clinical variables reflecting the sum of chronic comorbidity conditions for each household were included in the second step estimation; Charlson Comorbidity Index (CCI) and Principal In-Patient Diagnostic Cost Group (PIPDCG).[39] These two variables are similar in some sense but they are developed for the different purposes.[40] Also note that our CCI was not adjusted by age factor, since we have a separate AGE variable.

Table 6 shows the comparison between the estimation result from the panel data selection model and the results from the fixed effect model without correcting selection bias.[41] Even though the total charge is not the true cost but we believe that it is a good proxy for the true health resource utilization. Also notice that we did not take a logarithm on the total charge variable. This case is easier for the interpretation of estimation results and simpler for the formation of model. To take log transformation, we need to separate out zero total charges from the positive total charges by an additional nonparametric sample selection model.[42] Since a parametric sample selection model in panel data is sensitive for the choice of underlying distribution, whether multivariate logit or multivariate probit.[43] Also if the reason to take logarithm is to reduce noise as stated in Duan *et al*. (1983), panel data models include unobserved heterogeneity term in the model to capture whether they are really a noise or significant information (if extreme values are observed for an

[39] The conversion of diagnostic codes to PIPDCG is based on Health Care Financing Administration file ICD2DCG.XLS (http://www.hcfa.gov/stats/hmorates/aapccpg.htm).

[40] See Charlson *et al*. (1987) and Ash *et al*. (1989).

[41] For comparison purpose, all the results are based on the same set of variables, which are statistically significant at 5%.for the selection model for both females and males.

[42] Note that a two part model cannot be a choice for panel data. Since the positive expenditure population is varying over time and the zero expenditures are more likely non-observable responses (below deductible) rather than "genuine zeros" (See Jones, 2000).

[43] As explained earlier, these two distributions are not similar whereas they are pretty close except for the tail parts in univariate case. Therefore, even in a single year cross-sectional data, our data would generate quite different results for logit and probit specification, since there is a small proportion of zeros, 9% and 11% for 1996 and 1997, respectively.

individual repeatedly, that cannot be a noise). In the actual estimation of selection model, we used a sixth order bias reducing kernel (which showed the best performance in Kyriazidou's Monte Carlo simulation):[44]

$$K(v) = 1.5\exp(-v^2/2) + 0.1\exp(-v^2/18)(1/\sqrt{9}) - 0.6\exp(-v^2/8)(1/\sqrt{4}) \tag{11}$$

with an initial bandwidth $h_N = N^{-1/13}$ and $\delta = 0.1$ for correcting the asymptotic bias as in (10).[45]

The estimation results show the expected patterns. There are significant nonlinear effect of age on the total charge, which can be interpreted as older people use more health care resources but it is increasing with a diminishing rate. Hospitalization (Inpatient days) is a significant predictor of total charge. Comorbidity increases total charge but the way we build comorbidity condition (taking the sum of observed comorbidities of each household member) resulted insignificance of Coverage variable (a proxy for the number of household members).

All the three FE selection model results are based on the same $\hat{\beta}$ from the first step estimation. Therefore, we did not assume any different health plan type choice behavior by gender. From the comparison by gender, we can see the differences; age effect is much less for the females and the contribution of two comorbidity variables are much less for the females, and finally the effect of inpatient days are much bigger for the females. Therefore, we can conclude that there was a significant difference in health resource utilization by gender.

The usual FE model is based on the fixed effect estimator without correcting for the selection bias.

[44] The general multivariate formula to construct bias reducing kernel can be found in Bierens (1987). Also note that a higher order bias reducing kernel can have small negative values for a large argument (See Härdle, 1990, Chapter 4).

[45] For the sample including both females and males, the raw estimator ($\hat{\gamma}$) was 8429.65, -100.18, 1821.97, 218.78, 1164.51 for AGE, AGE^2, InDays, CCI, and PIPDCG, respectively. The corresponding asymptotic bias corrected estimator ($\tilde{\gamma}$) from Table 6 is 8333.47, -99.06, 1755.06, 209.54, 1155.02 for AGE, AGE^2, InDays, CCI, and PIPDCG, respectively. Note that the standard error is not changed by this asymptotic bias correction.

Therefore, the results of usual FE model reflects only the characteristics of PPO plan holders. When we compare the selection models with the usual FE models, the sign of coefficient estimates did not change but the magnitudes are somewhat different. The effect of Inpatient days and the effect of CCI are reduced by selection bias correction while the effect of age and PIPDCG are increased. The difference for the effect of age can be interpreted as the effect of age becomes more substantial for an average employee after considering the fact that the older people chose PPO plan (See Table 4). The other differences may be from the asymmetric influence of underlying factors like age on the health plan type choice.

5 Concluding Remarks

In this paper, we showed a selection bias correction for the medical expenditure estimation in panel data setup, based on the choice between HMO and PPO. The choice behavior was estimated with a high accuracy since our data was very stable and there was no systematic changes for the time periods. Also the results were consistent with the general perception of HMO plans and PPO plans. In other words,, HMO plan holders care more about medical costs while PPO plan holders care more about the quality of service. Since the PPO plans in our data set are offered by an Academic Health Center, which is generally believed to offer a better quality service in terms of technology, this comparison of price vs. quality makes more sense. The panel data methodology was especially useful for predicting health plan type choice because of the high correlation between the current choice and the past choice in our data.

The traditional issue of selection bias was discussed in terms of the different risk levels between different types of health care plans. When HMO's started to become popular, there was a concern that HMO's were accepting the lower risk group of people and as a result, an employer does not save medical cost (called

cream skimming). The Health Care Financing Administration (HCFA) has a similar problem to set a risk adjustment rates for Medicare managed care organizations. Since the Medicare data set (for example, Medicare Current Beneficiary Survey (MCBS)) has a large number of observations, it would be ideal to apply a nonparametric sample selection model like the one explained in this paper. Nonparametric modeling avoids any parametric specification so that it is less vulnerable to a misspecification problem, which is critical in applied researches. However, nonparametric models have slower rate of convergence to the normal distribution than parametric counter parts. To achieve a similar convergence rate, the nonparametric model needs much larger data set than the corresponding parametric model. Obviously, the natural panel structure of Medicare data set would appeal for the nonparametric panel data selection model.

For other applicable health data set, any data set has a panel data structure and a poor set of explanatory variables appeal for a panel data analysis methodology. The omitted individual specific (time invariant) explanatory variable can be controlled by either a fixed effect model or a random effect model. After controlling for these effects, the estimated coefficients for the available variables would have better predictive power than the popular cross-sectional approaches in health economics.

Even though we used a random effect model for the choice of health plan type, a fixed effect model can be applied with many new econometric techniques such as simulation based estimations (for example, Hajivassiliou and Ruud, 1994) or Manski's maximum score estimator (Manski, 1987), which is suggested in Kyriazidou's original paper. However, the choice between random effect modeling and fixed effect modeling in panel data analysis should be determined by a type of research question. The prevailing Hausman test to determine fixed effect against random effect is for testing specifications not modeling. If one finds a fixed effect model specification is meaningful by this test, she can also include time average of independent

variables in the random effect model to deal with possible correlation between unobserved heterogeneity parameter and independent variables (Chamberlain, 1980, 1984).

6 References

Abramovitz, M., Stegun, I., 1972. Handbook of Mathematical Functions. Dover Publications, New York.

Ahn, J., 2000. A Study of Employee Health Plan Choice and Medical Cost: Panel Data Probit Regression and Sample Selection Model. Ph.D. dissertation, University of Southern California.

Ai, C., Norton, E.C., 2000. Standard errors for the retransformation problem with heteroskedasticity. Journal of Health Economics 19, 697–718.

Altman, D., Cutler, D.M., Zeckhauser, R.J., 1998. Adverse selection and adverse retention. American Economic Review 88(2), 122–126.

Amemiya, T., 1981. Qualitative response models: a survey. Journal of Economic Literature 19(4), 483–536.

Ash, A., Porell, F., Gruenberg, L., Sawitz, E., Beiser, A., 1989. Adjusting Medicare capitation payments using prior hospitalization data. Health Care Financing Review 10(4), 17–29.

Baltagi, B.H., 1995. Econometric Analysis of Panel Data. John Wiley & Sons, West Sussex, U.K.

Barringer, M.W., Mitchell, O.S., 1994. Workers' preferences among company- provided health insurance plans. Industrial and Labor Relations Review 48, 141–152.

Ben-Porath, Y., 1973. Labor force participation rates and supply of labor. Journal of Political Economy 81, 697–704.

Bierens, H.J., 1987. Kernel estimators of regression functions, in: Bewley, T.F. (Ed.), Advances in Econometrics: Fifth World Congress, Vol. 1. Cambridge University Press, New York.

Buchmueller, T.C., 1995. Health risk and access to employer-provided health insurance. Inquiry 32, 75–86.

Buchmueller, T.C., Feldstein, P.J., 1997. The effect of price on switching among health plans. Journal of Health Economics 16, 231–247.

Butler, J., Moffit, R., 1982. A computationally efficient quadrature procedure for the one factor multinomial probit model. Econometrica 50, 761–764.

Chamberlain, G., 1980. Analysis of covariance with qualitative data. Review of Economic Studies 47, 225–238.

Chamberlain, G., 1984. Panel data, in: Griliches, Z., Intriligator, M. (Eds.), Handbook of Econometrics, Vol. 2. North-Holland, Amsterdam, pp. 1247–1318.

Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R., 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Disease 40(3), 373–383.

Cragg, J., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39, 829–844.

Dowd, B., Feldman, R., Moscovice, I., Wisner, C., Bland, P., Finch, M., 1996. An analysis of selectivity bias in the Medicare AAPCC. Health Care Financing Review 17(3), 35–57.

Duan, N., Smearing Estimate: A nonparametric retransformation method. Journal of American Statistical Association 78, 605–610.

Duan, N., Manning, W., Morris, C., Newhouse, J., 1983. A comparison of alternative models for the demand for medical care. Journal of Business & Economic Statistics 1, 115–126.

Duan, N., Manning, W., Morris, C., Newhouse, J., 1984. Choosing between the sample selection model and the multi-part model. Journal of Business & Economic Statistics 2, 283–289.

Duan, N., Manning, W., Morris, C., Newhouse, J., 1985. Comments on selectivity bias. Advances in Health Economics and Health Services Research 6. 19–24.

Eggers, P., 1980. Risk differential between Medicare beneficiaries enrolled and not enrolled in an HMO. Health Care Financing Review 2, 91–99.

Ellis, R.P., 1985. The effect of prior-year health expenditures on health coverage plan choice, in: Scheffler, R.M., Rossiter, L.F. (Eds.), Advances in Health Economics and Health Services Research, Vol. 6. JAI Press, Greenwich, CT, pp. 149–170.

Ettner, S.L., 1997. Adverse selection and the purchase of Medigap insurance by the elderly. Journal of Health Economics 16, 543–562.

Evans, W.N., Levy, H., Simon, K.I., 2000. Data watch: research data in health economics. Journal of Economic Perspectives 14(4), 203–216.

Feldman, R., Finch, M., Dowd, B., Cassou, S. 1989. The demand for employment-based health insurance plans. The Journal of Human Resources 24(1), 115–142.

Greene, W.H., 1997. Econometric Analysis, 3rd edn. Prentice Hall, Upper Saddle River, NJ.

Grootendorst, P.V., 1997. Health care policy evaluation using longitudinal insurance claims data: an application of the panel tobit estimator. Health Economics 6(4), 365–382.

Hajivassiliou, V.A., Ruud, P.A., 1994. Classical estimation methods for LDV models using simulation, in: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, Vol. 4. North-Holland, Amsterdam, pp. 2383–2441.

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, New York, NY.

Hay, J., Olsen, R.J., 1984. Let them eat cake: a note on comparing alternative models of the demand for medical care. Journal of Business and Economics Statistics 2, 279–282.

Hay, J., Leu, R., Rohrer, P., 1987. Ordinary least squares and sample-selection models of health-care demand. Journal of Business and Economics Statistics 5, 499–506.

Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5, 475–492.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47(1), 153–161.

Hellinger, F.J., 1995. Selection bias in HMO's and PPO's: a review of the evidence. Inquiry 32, 135–142.

Honoré, B., 1992. Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. Econometrica 60, 533–565.

Hornbrook, M.C., Bennett, M.D., Greenlick, M.R., 1989. Adjusting the AAPCC for selectivity and selection bias under Medicare risk contracts, in: Scheffler, R.M., Rossiter, L.F. (Eds.), Advances in Health Economics and Health Services Research, Vol. 10. JAI Press, Greenwich, CT, pp.111–149.

Horowitz, J., 1992. A smoothed maximum score estimator for binary response model. Econometrica 60, 505–531.

Hsiao, C., 1986. Analysis of Panel Data (1st ed.). Cambridge University Press, New York, NY.

Hsiao, C., 2001. Panel data models, in: Baltagi, B. (Ed.), Companion of Econometrics. Blackwell, Oxford.

Iezzoni, L. I., 1997. Risk Adjustment for Measuring Healthcare Outcomes. Health Administration Press, Chicago, IL.

Jones, A.M., 2000. Health econometrics, in: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics, Vol. 1A. North-Holland, Amsterdam, The Netherlands, pp. 265–344.

Kyriazidou, E., 1997. Estimation of a panel data sample selection model. Econometrica 65, 1335–1364.

Leung, S.F., Yu, S., 1996. On the choice between sample selection and two-part models. Journal of Econometrics 72, 197–229.

Maddala, G.S., 1985a. A survey of the literature on selectivity bias as it pertains to health care markets. Advances in Health Economics and Health Services Research 6, 3–18.

Maddala, G.S., 1985b. Further comments on selectivity bias. Advances in Health Economics and Health Services Research 6, 25–26.

Manning, W.G., 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics 17, 283–295.

Manning, W.G., Duan, N., Rogers, W., 1987. Monte Carlo evidence on the choice between sample selection and two part models. Journal of Econometrics 35, 59–82.

Manski, C., 1987. Semiparametric analysis of random effects linear models from binary panel data. Econometrica 55, 357–362.

Mullahy, J., 1998. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. Journal of Health Economics 17, 247–281.

Mundlak, Y., 1978. On the pooling of time series and cross section data. Econometrica 46, 69–85.

Neudeck, W., Podczeck, K., 1996. Adverse selection and regulation in health insurance markets. Journal of Health Economics 15, 387–408.

Riley, G., Tudor, C., Chiang, Y., Ingber, M., 1996. Health status of Medicare enrollees in HMO's and fee for service in 1994. Health Care Financing Review 17(4), 65–76.

Van de Ven, W.P.M.M., Ellis, R.P., 2000. Risk adjustment in competitive health plan markets, in: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics, Vol. 1A. North-Holland, Amsterdam, The Netherlands, pp. 755–845.

Van de Ven, W.P.M.M., van Vliet, R., 1995. Consumer information surplus and adverse selection in competitive health insurance markets: an empirical study. Journal of Health Economics 14, 149–169.

Table 1. Descriptive Statistics[46]

Health Plan Choice Estimation		Whole Data		Balanced Panel	
Variable	Description	'96Data (N=7743)	'97Data (N=7762)	'96Data (N=6644)	'97Data (N=6644)
D_{HMO}	Dummy variable for HMO	0.502	0.497	0.502	0.502
AGE	Age of employee	42.278 (11.640)	42.335 (11.636)	42.698 (11.353)	43.698 (11.353)
Working%	Working Status (full time = 100)	97.230 (11.560)	96.459 (12.704)	97.393 (11.373)	96.502 (12.728)
EX	Experience at USC (in months)	10.073 (8.280)	10.167 (8.362)	10.424 (8.221)	11.424 (8.221)
Coverage	Number of people covered (Single=1, Plus One=2, Family=3)	1.959 (0.864)	1.973 (0.867)	2.003 (0.863)	2.024 (0.861)
PREM	Monthly Premium	48.373 (37.975)	50.586 (35.803)	49.531 (38.436)	52.212 (36.980)
D_{FEMALE}	Dummy variable for female = 1	0.481	0.485	0.526	0.526
Total Charge Estimation (Data from PPO plan holders)		'96Data (N=3848)	'97Data (N=3902)	'96Data (N=3260)	'97Data (N=3260)
TCharge	Total Charge	9356.89 (38550.97)	8690.71 (39007.59)	9347.52 (36498.43)	9739.68 (39929.97)
InDays	Inpatient Days	0.78 (6.06)	0.81 (7.46)	0.76 (6.07)	0.91 (7.91)
CCI	Charlson Comorbidity Index	3.56 (16.87)	3.38 (20.08)	3.42 (14.21)	3.93 (21.88)
PIP-DCG	Principal Inpatient Diagnostic Cost Group (by the HCFA definition)	2.66 (6.28)	2.13 (5.96)	2.69 (6.29)	2.42 (6.32)
AGE	Same as above	43.88 (11.77)	43.93 (11.73)	44.44 (11.43)	45.44 (11.43)

[46] Mean is rounded at 1/1000 and standard Deviation is in the parentheses.

Table 2. Employee Choice of Health Plans

Whole Population	'97 Plans						
'96 Plans	Net_1	Net_2	Kaiser	Pacificare	CA_Care	Outside	Total
Net_1	3157	0	29	11	12	528	3737
Net_2	5	98	0	0	0	13	116
Kaiser	67	0	2470	11	7	383	2938
Pacificare	25	0	6	568	9	107	715
CA_Care	17	0	4	2	134	66	223
Cygna	1	0	3	6	2	2	14
Outside	533	3	387	86	109	3754	4872
Total	3272	98	2512	598	164	4853	12615

Note: 162 employee changed type (HMO vs. PPO) of plan (2.4%) and 217 employee changed their plan to different plan (3.3%) in the balanced panel of 6644 employees (excluding the employees associated with the outside options, 4872 and 4853 employees for 1996 and 1997, respectively). Also Cygna is completely discontinued from 1997.

Table 3. Change of Plan Coverage

	’97 Coverage				
’96 Coverage	Single	Plus One	Family	Outside	Total
Single	2319	120	25	592	3056
Plus One	61	1520	112	256	1949
Family	7	71	2409	251	2738
Outside	638	213	267	3754	4872
Total	3025	1924	2813	4853	12615

Table 4. Estimation Result of Random Effect Panel Probit Model (Gaussian Quadrature Approximation)[47]

Dependent Variable: D_{HMO}		
$N = 8861$		
Indepedent Variables	Estimate	p-value
Constant	-0.3601 (-2.21)	0.027
RPrem	-0.1979 (-26.09)	< 0.001
$RPrem^2$	0.0004 (25.25)	< 0.001
EX	0.0327 (6.09)	< 0.001
Coverage	0.9093 (11.66)	< 0.001
AGE	-0.0190 (-4.30)	< 0.001
ρ (randomeffect) $[S.D]$	0.6463 $[0.0321]$	See LR-test below
Log Likelihood	-1527.6561	
Wald test	$713.71 > \chi^2_5$	< 0.001
LR test: $\rho = 0$	$1677.98 > \chi^2_1$	< 0.001

[47] t-values (standard normal test can be used in a large sample) are in parentheses and standard error is in brackets for ρ. Minimum precision of p-values is at 1/1000. 30 quadrature points used for Gauss-Hermite quadrature approximation.

Table 5. Prediction Success Table

(Based on Panel Data Model from Table 4)

	Predicted Choice		observed count
Observed Choice	PPO	HMO	
PPO	7759	0	7759
HMO	50[a]	7696	7746
Predicted Count	7809	7696	15505

a.All fifty wrong predictions are from year 1996 sample. No wrong prediction at all for 1997 sample.

Table 6. Panel Selection Model Medical Expenditure Estimation (t-values are in parentheses)

A sixth order bias reducing kernel
$K(v) = 1.5\exp(-v^2/2) + 0.1\exp(-v^2/18)(1/\sqrt{9}) - 0.6\exp(-v^2/8)(1/\sqrt{4})$ was used along with two different bandwidth $h_N = N^{-1/13}$ and $h_N = N^{-0.1/13}$ for asymptotic bias correction as in (10).

Model	FE Selection Model			Usual FE Model		
Data	Female	Male	Both	Female	Male	Both
AGE	$6023.68^{\mathbf{s}}$ (2.87)	9573.87 (1.67)	$8333.47^{\mathbf{s}}$ (2.27)	2397.24 (1.08)	6789.03 (1.73)	4171.05 (1.77)
AGE^2	$-59.99^{\mathbf{s}}$ (-2.73)	-125.39 (-1.88)	$-99.06^{\mathbf{s}}$ (-1.98)	-33.84 (25.12)	-70.32 (-1.73)	-48.81 (-1.92)
InDays	$4387.96^{\mathbf{s}}$ (5.91)	$1503.05^{\mathbf{s}}$ (2.25)	$1755.06^{\mathbf{s}}$ (2.35)	$4754.12^{\mathbf{s}}$ (69.76)	$2718.93^{\mathbf{s}}$ (28.40)	$3478.16^{\mathbf{s}}$ (53.88)
CCI	27.20 (0.40)	330.59 (0.89)	$209.54^{\mathbf{s}}$ (1.99)	$288.46^{\mathbf{s}}$ (11.43)	$413.58^{\mathbf{s}}$ (6.24)	$389.94^{\mathbf{s}}$ (12.23)
PIPDCG	-195.31 (-0.70)	$1962.97^{\mathbf{s}}$ (2.02)	$1155.02^{\mathbf{s}}$ (2.28)	-43.21 (-0.58)	$694.89^{\mathbf{s}}$ (5.38)	$316.91^{\mathbf{s}}$ (3.97)
R^2	0.667	0.323	0.351	0.787	0.411	0.576
F-statistic	5961.50	4.48	13.21	1272.96	226.94	787.55

s: significant at 5% level

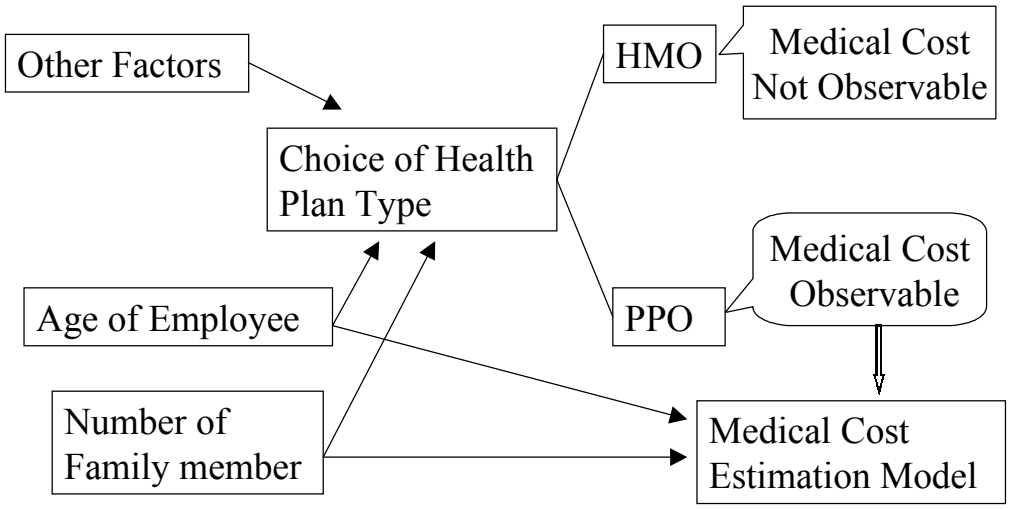

Figure 1: An Example Diagram of Medical Cost Estimation and Health Plan Type Choice

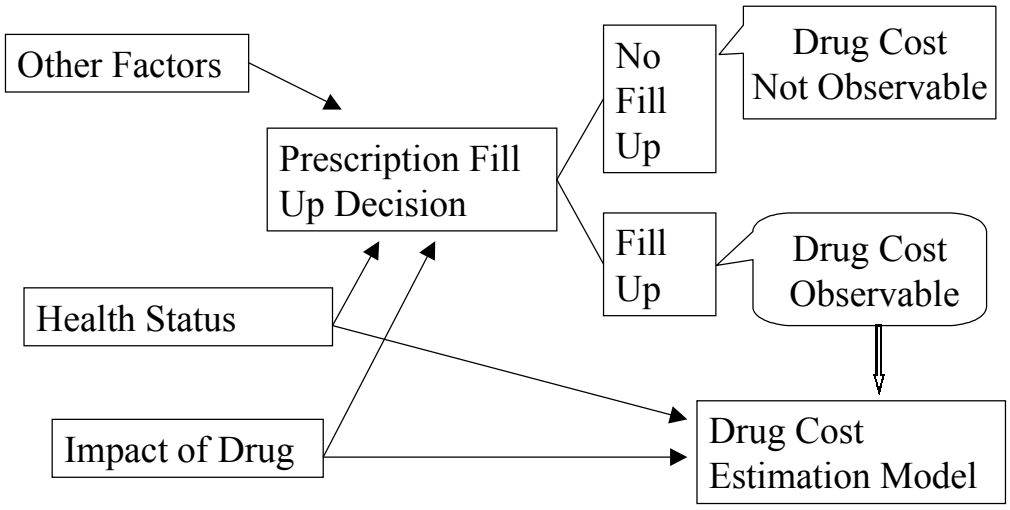

Figure 2: An Example Diagram of Prescription Fill Up Decision and Drug Cost Estimation