A Quasi-experimental Comparison of Econometric Models for Health Care Expenditures*

Partha Deb
Indiana University-Purdue University Indianapolis and
Hunter College, City University of New York

James F. Burgess, Jr.
US Department of Veterans Affairs Management Science Group and
Boston University School of Public Health

May 2003

Abstract

Individual health care expenditures have complex non-normal distributions with severe positive skewness and leptokurtosis. These features present severe challenges to reliable modeling of expenditures for prediction purposes. We compare a variety of methods using quasi-experimental techniques. Our quasi-experiments combine the distributional realism of actual data on health care expenditures with the reliability of Monte Carlo experimental results. We find that models based on Gamma densities predict substantially better than models based on linear regression with and without transformation of the dependent variable. Models based on finite mixtures of Gamma densities show further improvement in predictive properties.

*This research was partially funded by the US Department of Veterans Affairs, Management Science Group. We wish to thank Will Manning for helpful suggestions and participants at the 2002 Northeast Regional Health Economics Conference for their comments. Opinions and commentary offered are the responsibility of the authors and do not reflect official positions of the US Department of Veterans Affairs.

1 Introduction

A major concern of policymakers in the United States is the rapidly increasing cost of health care. Most cognoscenti believe that significant resources are spent on medical care that is not productive at the margin. Two important reasons for non-optimal use of resources are adverse selection into health insurance plans and moral hazard in the use of care. Although both these issues have been analyzed theoretically and empirically, only recently have practical attempts been made to reduce their effects using risk adjustment systems. Recently, some case studies describing the recent experiences of private sector employers with various risk adjustment systems have been published (e.g., Bertko et al., 1998, Dunn et al., 1998, Knutson et al., 1998).

We believe there are two important reasons for the relatively few attempts at designing and implementing risk adjustment systems. First, expected net plan revenues increase from adjusting payments based on enrollee risk when three market conditions are met (Frank and Rosenthal, 2001); plans must share in the costs of high expenditure patients, they must compete for enrollees, and they must believe that they are likely to attract an unfavorable mix of patients with high expenditures. Most plans face these conditions in very muted fashion. Consequently, market forces for risk adjustment of health care expenditures are relatively weak. Second, health care expenditures are so skewed, with rare high cost events, that most enrollees are 'healthier than average', a real life Lake Wobegon effect. This, combined with the fact that health care expenditures have a density not easily characterized by known parametric forms (Jones, 2001), makes predicting expenditures adequately a very difficult exercise.

Our research on modeling health care expenditures has a statistical fo-

cus, but it is motivated by the practical issues of building risk adjustment systems. It extends the literature in two important ways. First, we propose the use of finite mixture models for estimating health care expenditures which can serve as approximations to unknown probability densities (Lindsay, 1995; McLachlan and Peel, 1999). Second, we conduct an extensive evaluation of a number of econometric models in a quasi-experimental framework which combines the rigor of Monte Carlo experiments with the distributional realism of actual data, which is what the models must fit in practice. Although our paper focuses on the distribution of health care expenditures, the methods are readily applicable in other contexts where the variable of interest has complex distributional properties. The distribution of income is a leading example of such a variable (McDonald and Mantrala, 1995).

Earlier work on modeling individual health care expenditures focused on the use of transformations of the dependent variable in linear regression models to improve the quality of estimates and predictions. Recent research has considered generalized linear models for estimating expenditures. Blough, Madden and Hornbrook (1999) demonstrate the feasibility of such models but do not directly compare their models with standard approaches. Manning and Mullahy (2001) show that the generalized linear model based on the Gamma density has promise, but also that some classes of generalized linear models are considerably more sensitive to data problems than OLS. In general, known, parametric densities are inadequate approximations to the true densities for health care expenditures, and robust estimators typically sacrifice precision.

Finite mixture models are, in principle, semiparametric and can approximate any probability density. In practice, however, they should be viewed

as flexible extensions of parametric models, potentially providing a compromise between strongly parametric and fully semiparametric models. Finite mixture models provide a natural and intuitively attractive representation of heterogeneity that is clustered in a finite number of latent classes. The choice of the number of components in the mixture determines the number of classes, and the functional form for the density accommodates heterogeneity within each component. A consistent empirical finding is that distributions for unobservables can be approximated by low-dimensional finite mixtures. (Heckman, 2001).

Deb and Trivedi (1997, 2002) have demonstrated the superior performance of finite mixtures in modeling counts of health care utilization. Deb and Holmes (2001) show that a finite mixture model for positive mental health care expenditures provides more reliable estimates than does a log regression model. Consequently, we evaluate a class of finite mixture models for health care expenditures in our quasi-experimental approach.

We develop a quasi-experimental approach for evaluating models of unknown and complex data generating processes. As in Monte Carlo experiments, confidence in results is achieved through replication. However, our experimental samples are not drawn from known distributions. Such data are unlikely to capture all the relevant features of the empirical distribution of health care expenditures. Instead, we assume that all relevant features of the empirical distribution of health care expenditures are present in the very large dataset we use so that sampling from it is equivalent to sampling from the distribution of health care expenditures in the population. The random samples from this population also represent "ideal" enrollees in the sense that there is no selection into the plan. To the extent that these data mimic features of health care expenditures in other populations,

our quasi-experimental samples will be informative for models of health care expenditures in those populations.

In the following section of the paper, we formally present the competing models used in this paper and discuss model comparison strategies. The data are described in section 3 and empirical results in section 4. We conclude in section 5.

2 Methods

2.1 Econometric Models

Let y_i denote health care expenditures for person i and x_i denote the set of covariates including the intercept. We estimate the following econometric models.

A linear conditional mean model is estimated using OLS so that

$$\begin{aligned} \widehat{\beta} &= \arg\min \sum_{i=1}^{N} \{y_i - x_i\beta\}^2, \\ \widehat{y}_p &= x_p\widehat{\beta}, \end{aligned} \tag{1}$$

where $\widehat{y}_p$ denotes a conditional prediction. OLS with a linear mean has the desirable feature that it provides an unbiased predictor of health expenditures regardless of the distribution of the error term and the presence of heteroskedasticity. Nevertheless, given the extreme skewness of health care expenditures, it is possible that point forecasts obtained from this model may not be very precise. Note that this model is equivalent to the GLM model based on the normal density with linear link.

Two widely applied alternatives to the linear mean in the OLS context

use transformations of the dependent variable. In the log model,

$$\begin{aligned} \widehat{\beta} &= \arg\min \sum_{i=1}^{N} \{\log(y_i) - x_i\beta\}^2 , \\ \widehat{y_p} &= \exp(x_p\widehat{\beta}) \cdot \frac{1}{N} \sum_{i=1}^{N} \exp\left\{\log(y_i) - x_i\widehat{\beta}\right\}, \end{aligned} \tag{2}$$

and in the square root model,

$$\begin{aligned} \widehat{\beta} &= \arg\min \sum_{i=1}^{N} \{\sqrt{y_i} - x_i\beta\}^2 , \\ \widehat{y_p} &= (x_p\widehat{\beta})^2 + \frac{1}{N} \sum_{i=1}^{N} \left\{\sqrt{y} - x_i\widehat{\beta}\right\}^2, \end{aligned} \tag{3}$$

where the second term in each formula for the conditional prediction is a nonparametric smearing factor needed to retransform the prediction into the raw scale. Although these transformed models are designed to account for the skewness in health expenditures and the retransformation factors do not depend on normality of the errors, their predictions are not robust to heteroskedasticity in the transformed scale.

The model with the linear mean has an added advantage over models with complex mean specifications in that the regression coefficients are the average incremental costs of each disease and hence can be used to assess the face validity of the regressions. If used for rate setting, for example, plan managers would be very uncomfortable with negative regression coefficients or coefficient values outside the range of their intuitive expectations. So, often such models are recalibrated in ad hoc fashion until no "offending" coefficients remain (Ellis, R. P., personal communication). On the other hand, while the log and square root models generate positive conditional mean forecasts regardless of whether individual coefficients are positive or

negative, the linear mean model without such ad hoc adjustments may generate negative predictions. To assess the consequences of imposing such face validity, i.e., restricting the conditional mean to be positive, we use the estimates from the linear OLS model to generate predictions of the form

$$\widehat{y}_p = \max(x_p\widehat{\beta}, 0). \tag{4}$$

The second set of models are in the GLM class (McCullagh and Nelder, 1989). These models require only correct specification of the conditional mean for consistency and are quite flexible. We estimate GLMs based on the Gamma density as these have been shown to have desirable properties. We consider linear and squared mean specifications so that

$$\begin{aligned} \widehat{\beta} &= \arg\max \sum_{i=1}^{N} \left\{ -\frac{y_i}{x_i\beta} + \log\left(\frac{1}{x_i\beta}\right) \right\}, \\ \widehat{y}_p &= x_p\widehat{\beta} \end{aligned} \tag{5}$$

and

$$\begin{aligned} \widehat{\beta} &= \arg\max \sum_{i=1}^{N} \left\{ -\frac{y_i}{(x_i\beta)^2} + \log\left(\frac{1}{(x_i\beta)^2}\right) \right\}, \\ \widehat{y}_p &= (x_p\widehat{\beta})^2, \end{aligned} \tag{6}$$

respectively.

Finally, we estimate 2 models that are based on finite mixtures of densities. The random variable y_i in a finite mixture model is assumed to be a drawn from an additive mixture of C distinct subpopulations or components in proportions $\pi_1, ..., \pi_C$, where $\sum_{j=1}^{C} \pi_j = 1,\ \pi_j \geqslant 0\ (j = 1, ..., C)$. The mixture density for observation i, $i = 1, ..., n$, is given by

$$f(y_i|\boldsymbol{\theta}) = \sum_{j=1}^{C-1} \pi_j f_j(y_i|\boldsymbol{\theta}_j) + \pi_C f_C(y_i|\boldsymbol{\theta}_C), \quad i = 1, ..., n, \tag{7}$$

where $\pi_C = 1 - \sum_{j=1}^{C-1} \pi_j$. Each term in the sum on the right-hand side is the product of the mixing probability π_j and the component density $f_j(y_i|\boldsymbol{\theta}_j)$ which has parameters $\boldsymbol{\theta}_j$. In general, the π_j are unknown and estimated along with $\boldsymbol{\theta}_j$. A labelling restriction that $\pi_1 \geq \pi_2 \geq .... \geq \pi_C$, which can always be satisfied by rearrangement, is required for identification (normalization). Given our success with the gamma density in preliminary analysis, we consider models based on mixtures of gamma's:

$$\begin{aligned} \widehat{\beta}_j, \widehat{\pi}_j &= \arg\max \sum_{i=1}^{N} \log \left\{ \sum_{j=1}^{C-1} \pi_j \cdot \exp\left(-\frac{y_i}{x_i\beta_j}\right)\left(\frac{1}{x_i\beta_j}\right) \right\}, \qquad (8) \\ \widehat{y}_p &= \sum_{j=1}^{C-1} \widehat{\pi}_j x_f \widehat{\beta}_j, \quad j = 1, 2, ..., C. \end{aligned}$$

where β_j and π_j are estimated jointly.

We consider finite mixture models with linear mean specifications and two or three gamma component densities. Although both the specification of the mean and the number of components are trivially modified in principle, we restrict our attention to linear mean specifications for reasons of face validity discussed above and to two and three components for computational feasibility given the large scale of our study. Note that the model given by (8) is a generalization of (5); however, the two- and three-component mixture models could possibly perform worse than their one-component (degenerate) counterparts in finite samples.

Table 1 provides labels for each of the models considered in our experiments along with brief descriptions of the estimation method and prediction functions. The labels are subsequently used in our description of the results.

2.2 Experimental Design

The study design is quasi-experimental. Monte Carlo principles are used to create 'experimental' samples and confidence in results is achieved through experimental replication. However, unlike 'true' Monte Carlo experiments, our 'experimental' samples are drawn from real data with unknown distribution rather than artificial data drawn from a known distribution. The complex data generating process for health care expenditures is well known not to follow any known parametric distribution and that characteristics of extreme observations make predicting health care expenditures a difficult exercise. Therefore, if we used data drawn from a known distribution in our study, it would likely not capture all the features of the empirical distribution of health care expenditures, and would have the additional drawback that it would always be possible to include an econometric model in the study that would *a priori* be closer to to the true data generating density (or even be correctly specified). Instead, we assume that all relevant features of the empirical distribution of health care expenditures are present in a very large dataset we use so that sampling from it is equivalent to sampling from the distribution of health care expenditures in the population. To the extent that these data mimic features of health care expenditures in other populations, our quasi-experimental samples will be informative for models of health care expenditures in those populations.

The dataset comes from FY2000 expenditures by users of the US Department of Veterans Affairs health care system. The 2,979,760 observations were randomly split into two groups: 1,500,000 observations were assigned to the estimation group and 1,000,000 to the prediction group. Note that these sub-groups themselves are quite large and reasonably might be treated

as pseudo-populations. We were restricted to these sizes by computer memory considerations. Samples of size $N \in \{10{,}000\ 50{,}000\ 100{,}000\ 200{,}000\ 500{,}000\}$ were drawn from the estimation group using simple random sampling with replacement. Note that most datasets from public or private populations in managed care plans or health care provider systems in the US fall in the range of our sample sizes (see e.g., Dunn, 1998, who analyzes risk adjustments in four samples of 70,000, 115,000, 120,000 and 240,000). The parameters of the models described above were estimated for each sample and saved. This process was repeated 20 times for each sample size. The parameters obtained from each replication were used to calculate conditional means using all million observations from the prediction group.

Two statistics were calculated to evaluate the quality of the predictions: the mean prediction error

$$MPE = \frac{1}{N_f} \sum_{i=1}^{N_f} \left(\widehat{y}_f - y_i \right), \tag{9}$$

and the mean absolute prediction error

$$MAPE = \frac{1}{N_f} \sum_{i=1}^{N_f} \left| \widehat{y}_f - y_i \right|. \tag{10}$$

MPE is a measure of predictive accuracy on average (across observations) while $MAPE$ is a measure of predictive accuracy for individual observations.

We also calculated each of these statistics after *trimming* the prediction sample by eliminating 0.5% of the largest expenditures ($N_f = 995,000$). We did this for two reasons. First, each of these statistics may be unduly affected by a very small fraction of extremely large expenditures in the prediction sample and these extreme observations may not regularly appear in smaller populations. Second, the design of many pricing schemes include reinsurance

for enrollees or patients with very large expenditures so models should be evaluated on the observations not eligible for reinsurance.

2.3 Response Surfaces

For any statistic of interest, ideally one would like to compute analytic formulae for its predicted values as a function of experimental characteristics. For example, if the statistic of interest is bias and one is interested in determining how bias decreases as the sample size increases, the ideal would be an analytic formula that relates bias to sample size. If these formulae are not known, it is possible to approximate them using polynomial approximations to the true functional forms. Regressions of these polynomial approximations are called response surfaces. Maasoumi and Phillips (1982), Hendry (1982), and Davidson and MacKinnon (1993) have detailed discussions of the merits of response surface methodology. In our context, one desirable feature of response surface methodology is that it facilitates understanding of experimental evidence because large amounts of experimental data can be summarized using simple functional forms. It also provides applied researchers a simple tool for computing outcomes at points in the design space that are not included in the experimental study. Another advantage, especially for computationally intensive processes, is that a large number of replications is not required. Each of these advantages of response surface methodology is important in the context of our study relative to simple tabulation of the results: we have many design points (model×sample size), interest in performance at other sample sizes, and very computationally intensive estimation.

Let the models in this study be numbered by $m = 1, 2, ..., 8$, let $s = 1, 2, ..., 5$ denote the different sample sizes and allow $r = 1, 2, ..., 20$ replica-

tions at each experimental design point. Let $d[m]$ denote dummy variables indicating the model on the basis of which the statistic, MPE or $MAPE$, was calculated. Let N_s denote the sample size used for estimation of the model. The response surfaces for MPE are specified as

$$MPE_{msr} = \sum_{m=1}^{8} \alpha[m]d[m] + \sum_{m=1}^{8} \frac{\gamma[m]d[m]}{N_s} + u_{msr}, \tag{11}$$

where $\alpha[m]$ and $\gamma[m]$ are regression coefficients. The second term in the right hand side of the regression reflects the fact that MPE is expected to decline at the rate N_s. Note that in each response surface regression, $\alpha[m]$ denotes the asymptotic expected value of MPE for model m. Expected MPE for desired finite sample sizes can be calculated by plugging in those sample sizes.

$MAPE$ only takes positive values, so its response surface is specified in logarithms, i.e.,

$$\log(MAPE_{msr}) = \sum_{m=1}^{8} \alpha[m]d[m] + \sum_{m=1}^{8} \frac{\gamma[m]d[m]}{N_s} + u_{msr}. \tag{12}$$

Now differences in values of $\alpha[m]$ represent percentage differences in $MAPE$ across models.

The regression specification for MPE evaluates models in terms of their ability, on average, to predict *average* expenditures in large samples. The regression specification for $MAPE$ evaluates models in terms of their ability, on average, to predict *individual* expenditures. It is possible that a model which predicts average expenditures well, on average across replication samples, may not predict average expenditures well in any particular replication sample, i.e., the dispersion of the distribution of MPE is also an important evaluation criterion. To address this issue, we estimate a third regression for

the logarithm of absolute deviations of MPE, denoted $ADMPE$, specified as

$$\begin{aligned} \log(ADMPE) &= \log\left|MPE_{msr} - \overline{MPE}_{ms.}\right| \\ &= \sum_{m=1}^{8} \alpha[m]d[m] + \sum_{m=1}^{8} \frac{\gamma[m]d[m]}{Ns} + u_{msr}, \end{aligned} \tag{13}$$

where

$$\overline{MPE}_{ms.} = \frac{1}{20}\sum_{r=1}^{20} MPE_{msr} \tag{14}$$

is the average MPE across replications within an experimental design point.

3 Data

The US Department of Veterans Affairs (VA) operates the largest health care system in the US with 163 hospitals, more than 800 community and facility-based clinics, 135 nursing homes, and other facilities. With a medical care budget of more than $19 billion in FY2000, VA provided care to 3.8 million unique users, 3,000,499 of whom were provided care under priority for service connected disabilities, meeting an income/wealth based means test, or from a variety of smaller health care need and veteran specific reasons and thus had full access to the health care services offered. 2,979,760 of these patients have measured costs accurate enough to be included in the patient sample that serves as the sampling population for our analysis as described above.

In recent years, the most important advances in risk adjusting patient populations to explain health care expenditures have employed diagnostic information to characterize disease patterns. There are two basic strands of analysis that flow from this work. Most commonly, analysis has focused on predicting the health care utilization of enrolled patient populations next

year from diagnoses and other information (possibly even including costs) collected this year. This is called prospective modeling. However, a growing application of risk adjustment is in helping integrated health care delivery systems or insurers understand differences in the risk of current populations, for budget allocation or rate setting purposes. We employ this type of concurrent modeling in this paper.

We provide summary statistics for costs in Table 2. The estimates are based on a sample size of 2,500,000 that comprise the combination of our estimation and prediction samples. As is well known for health care expenditures in other contexts, expenditures for the VA population are also highly skewed and leptokurtic. When logarithms of health care costs are examined, skewness and kurtosis are considerably smaller but still statistically significant. As a comparison, we also report summary statistics of health care expenditures for a representative sample of the US population in 1996 obtained from the Medical Expenditure Panel Survey (MEPS) and for the sub-sample of MEPS respondents enrolled in Medicare. The results in Table 2 show that the statistical characteristics of health care expenditures of the Medicare population are very similar to those of the VA population and that the distribution of health care expenditures for the US population overall are considerably more skewed and leptokurtic than either of the sub-populations. Note that as the data are refined into more homogeneous populations, the skewness and kurtosis moment measures fall.

To characterize the explainable portion of variation in expenditures, we employ Diagnostic Cost Group (DCG)/Hierarchical Coexisting Conditions (HCC) models (Ellis, et al. (1996), Ash, et al. (1998), Pope, et al. (1998)) to group ICD-9-CM diagnoses into HCC indicator groups as explanatory variables for health care expenditures. This model takes the 15,000 ICD-9-

CM codes, groups them into categories and then places the groups into body system/clinical condition specific hierarchies. These hierarchies allow some multiple HCC's and disallow others, helping to address overfitting problems when people with complex diagnoses also by definition have less complex ones in the same hierarchy. Out of the 118 HCC's in Version 5 of the DCG model, we employ 42 HCC's in our model that appear with a frequency of at least 1 percent in the sample of 2,500,000. Brief descriptions of the HCC's and their sample frequencies are reported in Table 3.

4 Results

As described above, 5 different sample sizes were considered for estimation and each experiment was replicated 20 times. OLS estimates are trivially obtained. The log likelihood functions of GLM models with gamma baseline density are typically well behaved so ML estimation is easy to conduct, though obviously computationally intensive for some of the larger sample sizes. The log likelihood functions of finite mixture models are not so well behaved in principle. They can have multiple optima. In practice, one can overcome this potential problem simply by experimentating with starting values, although more complex algorithms which attempt to avoid convergence to local optima are also available. But in this experimental setting, it was not feasible to ensure convergence to the global maximum in each case. For the two-component mixture model, we used starting values based on the converged estimates of the Gamma model (degenerate mixture) which it generalized. For the three-component mixture model, we used starting values based on the converged estimates of the two-component mixture. Although these are reasonable starting values, convergence to a local optimum

cannot be ruled out. Therefore, the results for the finite mixture models may be contaminated by non-maximized estimates, and thus should be treated as the worst case scenarios.

The experimental samples of MPE and $MAPE$ consist of 800 observations each. Response surface regressions specified by equations (11) − (13) were estimated and the parameter estimates are reported in Table 4. In each case in the regressions of MPE and $MAPE$, the R^2's are over 0.99. For the deviations of MPE regressions, the R^2's are over 0.90. Overall, the response surfaces are very well specified and capture most of the variation across experimental design points.

The asymptotic expected values of MPE indicate how the average value of predicted health care expenditures from a particular model compares to the average health care expenditure in the prediction sample. Linear and square root regression models estimated by OLS have negligible bias when evaluated using all observations in the prediction sample. All other models have substantially larger biases. Predictions from both finite mixture models are downward biased, but *FM3*-Γ-*linear* has a lower bias than *FM2*-Γ-*linear*. Once the prediction sample is trimmed, the finite mixture models [*FM2*-Γ-*linear* and *FM3*-Γ-*linear*] have the smallest biases of all models. *FM2*-Γ-*linear* is clearly the best model: it is virtually unbiased. Both *ols-linear* and *ols-square* root are upward biased. These results taken together indicate that the lower bias of the linear OLS model is due to its ability to predict the largest expenditures well. Note also that the OLS model with non-negative predictions (*ols-linear>0*) has a significant bias in each case and it overpredicts by about $60 relative to the standard OLS model (*ols-linear*) unless the prediction sample is trimmed.

As discussed above, the choice of functional form for the specification of

the conditional mean is important for a variety of reasons. Although the linear conditional mean has virtues in its simplicity and ease of interpretation, the squared and exponential conditional means have other virtues. The results show that the log regression model performs surprisingly poorly vis-a-vis the alternatives. It produces substantially upward biased predictions. In preliminary work we found that the gamma model with exponential link also performed very poorly hence was eliminated from further consideration. We have chosen to include the log regression model because it is a leading model among those used in existing empirical studies. Within the family of linear regression models, the linear and square root models have very similar MPE's. It is not possible to discriminate between the two models on this basis. In the case of GLM models based on the Gamma density, there are differences in MPE's between linear and squared links, but neither dominates.

The asymptotic expected values of $\log(MAPE)$ indicate how values of individual predicted health care expenditures from a particular model compare to the values of actual health care expenditures in the prediction sample. Models with lower values of $\log(MAPE)$ predict individual expenditures better than models with higher values. The results in Table 4 show that the two-component finite mixture model with gamma densities dominates the rest by the $MAPE$ criterion regardless of whether the prediction sample is trimmed or not. In the untrimmed case, *FM2*-Γ-*linear* has an 11 percentage point lower $MAPE$ than the linear regression model and 2 percentage point lower $MAPE$ than Γ-*linear*. Interestingly, the $MAPE$ of *FM3*-Γ-*linear* is worse than *FM2*-Γ-*linear*. There are two potential reasons for this decline in performance. First, it is possible that this is manifestation of a $MPE - MAPE$ trade-off which appears is many statistical contexts. In

our context, it would most likely be due to the fact that by increasing the ability of the model to predict the small number of high expenditures well, the three-component model was doing worse at predicting the large number of lower expenditures. Second, it is possible that the parameter estimates of three-component model are based, in a substantial fraction of cases, on log likelihood values that are not globally maximized. This is plausible because the log likelihood function of the three-component model has multiple optima, in principle. Unfortunately, a closer investigation of these possibilities is beyond the scope of this paper.

The OLS model with logarithmic transformation continues to perform very poorly. The $MAPE$ from the square root model is always lower than the $MAPE$ from the linear model in the regression case, but the relative performance of the linear and squared conditional means are reversed in the Gamma GLM models.

The regressions for $ADMPE$ evaluate models in terms of the dispersion in the distribution of MPE across replication samples. The results, reported in the last two columns of Table 4, show that, asymptotically, *ols-square root*, Γ-*linear* and *FM2*-Γ-*linear* have the smallest deviations. The simple linear regression model (*ols-linear*) is not far behind. Once again, models from the Gamma family are among the top performers.

The finite sample values of MPE, $MAPE$ and $ADMPE$, and their rates of convergence to the asymptotic values, also are described by the estimates in Table 4, but these are not transparent. Therefore, in Figure 1, we plot the expected values of MPE, $\log(MAPE)$, $\log(ADMPE)$ against estimation-sample size using trimmed and untrimmed prediction samples for three of the leading models with linear mean specifications - *ols-linear*, Γ-*linear* and *FM2*-Γ-*linear*. As was evident from the regression estimates,

ols-linear has the lowest bias when prediction samples is untrimmed, but *FM2*-Γ-*linear* has smaller bias when the prediction sample is trimmed. The rates of convergence of *ols-linear* and Γ-*linear* to the asymptotic MPE are very quick; 15,000-20,000 observations appear to be sufficient. On the other hand, convergence is slower for the finite mixture model, as expected. But even for *FM2*-Γ-*linear*, sample sizes of 30,000-40,000 are sufficient to ensure asymptotic values of MPE.

$MAPE$ converges at similar rates. The advantages of the models based on the gamma density vis-a-vis *ols-linear* are dramatic. The gains from using *FM2*-Γ-*linear* over *ols-linear* is in the order of 10-15 percentage points when estimation samples 20,000 or more observations are available. In the context of the budgets at stake in many rate-setting exercises, these gains are substantial.

Finally, $DAMPE$ is smallest for *FM2*-Γ-*linear*. Along this dimension, however, the three leading models are much more similar than along the MPE or $MAPE$ dimensions.

5 Conclusion

Many health outcome variables in health economics deviate from known parametric densities even upon transformation and reliable estimation methods for practical purposes continues to be an unsettled issue. The quasi-Monte Carlo experiments reported in this paper subject a number of plausible econometric models to tests in a variety of dimensions. The results demonstrate that models with linear mean specifications perform at least as well as models with more complex means or those that require retransformation. Linear regression models estimated by ordinary least squares

produce unbiased predictions, but individual predictions are relatively imprecise. Unbiased predictions in themselves are an inadequate criterion for a good econometric model because there exist incentives for providers to miscode, misreport, etc. when payments deviate substantially from costs. The ideal standard involves predicting individual expenditures as well as possible given the data. A GLM model based on the Gamma density with a linear link has reasonable bias properties and superior individual predictions vis-a-vis the linear regression model. A finite mixture model constructed using two Gamma densities with linear means has lower bias when the largest half percent of observations are removed from the prediction sample. The two-component finite mixture model has superior individual predictions in both trimmed and untrimmed prediction samples relative to the GLM model based on the Gamma density. Adding a third component to the mixture model appears to reduce biases but at the cost of poorer individual predictions.

As Frank and Rosenthal (2001) suggest, there are other options that health plans and other entities potentially interested in predictive risk adjustment have to attempt to reduce adverse selection. Such selection can be favorable, in trying attract systematically overpredicted groups, or unfavorable, in avoiding systematically underpredicted groups. To the extent that market and regulatory barriers to use of risk adjustment models continue to decline in the future, more and more attention will be paid to the need to accurately model these complex distributions. Careful analytic and theoretical work comparing problems of bias with problems of predicting individual expenditures would help resolve the need to develop appropriate loss functions to optimize particular strategies for risk adjustment.

In practice, estimation of finite mixture models raises some computa-

tional difficulties, especially as the number of points of support in the mixture distribution increases. This paper shows that there are substantial gains in predictive performance associated with the use of finite mixture models with two components, The finding is consistent with conventional wisdom and empirical evidence in the literature on finite mixture models that two to four points of support are typically sufficient. A small number of components is more likely to be sufficient if one starts with a baseline density that forms a reasonable first approximation to the true data density. Therefore, given the advances in computer hardware and statistical computing technology, the computational burden of finite mixture models should not discourage their use.

References

Ash A., R. P. Ellis, W. Yu, et al. (1998), "Risk Adjustment for the Non-Elderly." Final Report submitted to the Health Care Financing Administration under cooperative agreement No.18-C-90462/1-02, Boston, MA: Boston University, June 1998.

Bertko J. and , and S. Hunt (1998), "Case Study: The Health Insurance Plan of California", *Inquiry*, 32, 148-153.

Blough, D.K., C.W. Madden and M.C. Hornbrook (1999), "Modeling Risk using Generalized Linear Models, *Journal of Health Economics*, 18, 153-171.

Davidson, R. and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics*, Oxford University Press.

Deb, P. and A. M. Holmes (2000), "Estimates of Use and Costs of Behavioral Health Care: A Comparison of Standard and Finite Mixture Models", *Health Economics*, 9, 475-489.

Deb, P., and P.K. Trivedi (1997), "Demand for Medical Care by the Elderly in the United States: A Finite Mixture Approach", *Journal of Applied Econometrics*, 12, 313-336.

Deb, P., and P.K. Trivedi (2002), "The Structure of Demand for Health Care: Latent Class versus Two-part Models", *Journal of Health Economics*, 21, 601-625.

Dunn, D. (1998), "Applications of Health Risk Adjustments: What Can Be Learned to Date?" *Inquiry*, 35, 132-147.

Ellis, R.P., G. Pope, L.I. Iezzoni, J.Z. Ayanian, D.W. Bates, H. Burstin, and A. Ash (1996), "Diagnosis-Based Risk Adjustment for Medicare Capitation Payments." *Health Care Financing Review*, 17, 101-128.

Frank, R.G., and M.B. Rosenthal (2001), "Health Plans and Selection: Formal Risk Adjustment vs. Market Design and Contracts", *Inquiry*, 38, 290-298.

Heckman, J.J. (2001), "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture", *Journal of Political Economy*, 109, 673-748.

Hendry, D. F. (1982), "A Reply to Professors Maasoumi and Phillips", *Journal of Econometrics*, 19, 203-213.

Jones, A. M. (1999), "Health Econometrics", forthcoming in J. P. Newhouse and A. J. Culyer, eds., *Handbook of Health Economics*, Amsterdam: North Holland.

Knutson D. (1998), "Case Study: The Minneapolis Buyers Health Care Action Group", *Inquiry,* 32, 171-177.

Lindsay, B. J. (1995), *Mixture Models: Theory, Geometry, and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, IMS-ASA.

Maasoumi, E. and P. C. B. Phillips (1982), "On the Behavior of Inconsistent Instrumental Variable Estimators", *Journal of Econometrics*, 19, 183-201.

Manning, W.G. and J. Mullahy (2001), "Estimating Log Models: To Transform or Not to Transform?", *Journal of Health Economics*, 20, 461-494.

McDonald, J. B. and A. Mantrala (1995), "The Distribution of Personal Income: Revisited", *Journal of Applied Econometrics*, 10, 201-204.

McCullagh, P. and J.A Nelder (1989), *Generalized Linear Models*, London: Chapman and Hall.

McLachlan, G.J., and D. Peel (2000), *Finite Mixture Models,* New York: John Wiley.

Pope, G., R.P. Ellis, C. Liu, et al. (1998), "Revised Diagnostic Cost Group (DCG)/Hierarchical Coexisting Conditions (HCC) Models for Medicare Risk Adjustment." Final Report to the Health Care Financing Administration under Contract No. 500-95-048, Waltham, MA: Health Economics Research, Inc., February 1998.

Table 1
Description of Models

	Label	Estimation method	Prediction function
1	ols-linear	OLS	$x_i\beta$
2	ols-log	OLS	$\exp(x_p\widehat{\beta}) \cdot \frac{1}{N}\sum_{i=1}^{N} \exp\left\{\log(y_i) - x_i\widehat{\beta}\right\}$
3	ols-square root	OLS	$(x_p\widehat{\beta})^2 + \frac{1}{N}\sum_{i=1}^{N}\left\{\sqrt{y} - x_i\widehat{\beta}\right\}^2$
4	ols-linear>0	OLS	$\max(x_i\beta_j, \varepsilon)$
5	Γ-linear	ML, Γ density	$x_i\beta$
6	Γ-square	ML, Γ density	$(x_i\beta)^2$
7	FM2-Γ-linear	ML, mixture of 2 Γ's	$\sum_{j=1}^{C-1}\widehat{\pi}_j x_f\widehat{\beta}_j, \quad j = 1, 2$
8	FM3-Γ-linear	ML, mixture of 3 Γ's	$\sum_{j=1}^{C-1}\widehat{\pi}_j x_f\widehat{\beta}_j, \quad j = 1, 2, 3$

Table 2
Summary Statistics of Costs

	Cost/$1000			log(Cost/$1000)		
	VA	MEPS		VA	MEPS	
		All	Medicare		All	Medicare
N	2500000	18490	2588	2500000	18490	2588
Mean	5.342	2.372	6.185	0.411	-0.564	0.738
Median	1.537	0.527	2.097	0.430	-0.641	0.741
Std Deviation	14.804	8.572	12.521	-0.102	1.637	1.549
Skewness	9.717	21.287	6.388	-0.102	0.212	-0.201
Kurtosis	203.512	850.131	68.772	0.697	-0.151	0.161
99th percentile	70.322	29.852	58.440	4.253	3.396	4.068
95th percentile	22.612	9.491	25.773	3.118	2.250	3.249
75th percentile	3.839	1.699	5.883	1.345	0.530	1.772
25th percentile	0.586	0.180	0.798	-0.534	-1.715	-0.225
5th percentile	0.107	0.180	0.149	-2.235	-3.147	-1.904
1st percentile	0.032	0.180	0.038	-3.442	-4.605	-3.270

Table 3
Description of Covariates

Variable	Description	Frequency
HCC004	Other Infectious Disease	0.132
HCC006	High Cost Cancer	0.011
HCC007	Moderate Cost Cancer	0.012
HCC008	Low Cost Cancers/Tumors	0.050
HCC013	Diabetes with Chronic Complications	0.037
HCC014	Diabetes with Acute Complications	0.018
HCC015	Diabetes with No or Unspecified Complications	0.128
HCC017	Moderate Cost Endo/Metab/Fluid-Electrolyte	0.026
HCC020	High Cost Chronic Gastrointestinal	0.010
HCC022	Moderate Cost Gastrointestinal	0.048
HCC023	Low Cost Gastrointestinal	0.177
HCC025	Rheumatoid Arthritis/Connective Tissue	0.018
HCC028	Blood/Immune Disorders	0.013
HCC029	Iron Deficiency and Other Anemias	0.048
HCC030	Dementia	0.028
HCC031	Drug/Alcohol Dependence/Psychoses	0.065
HCC032	Psychosis/Higher Cost Mental	0.088
HCC033	Depression/Moderate Cost Mental	0.070
HCC042	High Cost Neurological	0.022
HCC043	Moderate Cost Neurological	0.049
HCC044	Low Cost Neurological	0.042
HCC048	Congestive Heart Failure	0.059
HCC049	Heart Arrhythmia	0.043
HCC051	Other Acute Ischemic Heart Disease	0.011
HCC053	Valvular and Rheumatic Heart Disease	0.022
HCC058	High Cost Cerebrovascular Disease	0.012
HCC059	Low Cost Cerebrovascular Disease	0.043
HCC060	High Cost Vascular Disease	0.052
HCC063	Other Circulatory Disease	0.024
HCC064	Chronic Obstructive Pulmonary Disease	0.118

Table 3 (continued)
Description of Covariates

Variable	Description	Frequency
HCC067	Low Cost Pneumonia	0.016
HCC075	Low Cost Ear, Nose, and Throat	0.184
HCC078	Renal Failure	0.020
HCC080	Other Urinary System	0.070
HCC091	Chronic Ulcer of Skin	0.017
HCC097	Other Injuries and Poisonings	0.111
HCC098	Complications of Care	0.017
HCC099	Major Symptoms	0.156
HCC100	Minor Symptoms, Signs, Findings	0.323
HCC113	Elective/Aftercare	0.129
HCC116	Rehabilitation	0.036
HCC118	History of Disease	0.063

Table 4
Response Surface Regressions

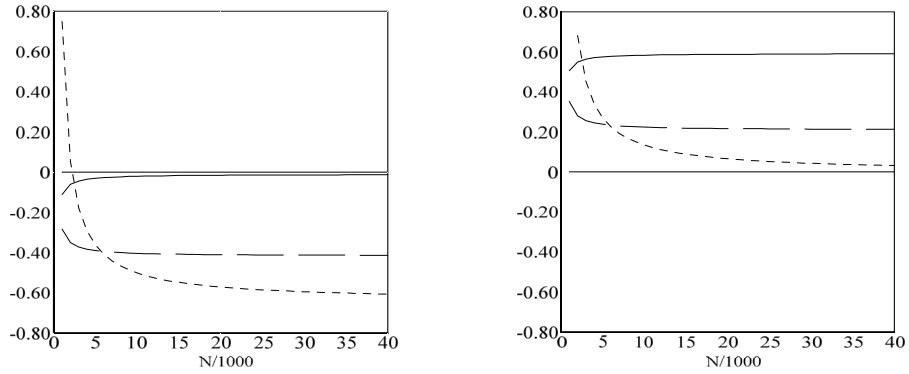
Dependent Variable	MPE		$\log(MAPE)$		$\log(ADMPE)$	
Prediction Sample	Untrimmed	Trimmed	Untrimmed	Trimmed	Untrimmed	Trimmed
α[ols-linear]	-0.010	0.592	1.516	1.380	-4.042	-4.094
	(0.014)	(0.013)	(0.002)	(0.002)	(0.147)	(0.144)
α[ols-log]	3.932	3.965	2.038	1.908	-3.029	-3.147
	(0.014)	(0.013)	(0.002)	(0.002)	(0.147)	(0.144)
α[ols-square root]	-0.013	0.604	1.463	1.314	-4.178	-4.170
	(0.014)	(0.013)	(0.002)	(0.002)	(0.147)	(0.144)
α[ols-linear>0]	0.050	0.653	1.502	1.364	-4.089	-4.145
	(0.014)	(0.013)	(0.002)	(0.002)	(0.147)	(0.144)
α[Γ-linear]	-0.417	0.210	1.431	1.273	-4.107	-4.132
	(0.014)	(0.013)	(0.002)	(0.002)	(0.147)	(0.144)
α[Γ-square]	0.138	0.716	1.448	1.307	-3.958	-4.012
	(0.014)	(0.013)	(0.002)	(0.002)	(0.147)	(0.144)
α[FM2-Γ-linear]	-0.641	-0.002	1.409	1.244	-4.150	-4.154
	(0.014)	(0.013)	(0.002)	(0.002)	(0.147)	(0.144)
α[FM3-Γ-linear]	-0.411	0.217	1.431	1.274	-3.043	-3.099
	(0.014)	(0.013)	(0.002)	(0.002)	(0.147)	(0.144)
γ[ols-linear]	-0.102	-0.087	0.156	0.175	11.540	10.852
	(0.312)	(0.281)	(0.035)	(0.039)	(3.210)	(3.131)
γ[ols-log]	1.635	1.367	0.195	0.189	19.164	17.902
	(0.312)	(0.281)	(0.035)	(0.039)	(3.210)	(3.131)
γ[ols-square root]	-0.027	-0.019	0.036	0.039	14.549	13.820
	(0.312)	(0.281)	(0.035)	(0.039)	(3.210)	(3.131)
γ[ols-linear>0]	-0.061	-0.046	0.151	0.169	11.383	12.121
	(0.312)	(0.281)	(0.035)	(0.039)	(3.210)	(3.131)
γ[Γ-linear]	0.134	0.142	0.094	0.108	13.928	13.494
	(0.312)	(0.281)	(0.035)	(0.039)	(3.210)	(3.131)
γ[Γ-square]	0.130	0.150	0.107	0.118	18.113	18.233
	(0.312)	(0.281)	(0.035)	(0.039)	(3.210)	(3.131)
γ[FM2-Γ-linear]	1.389	1.368	0.195	0.239	13.262	12.122
	(0.312)	(0.281)	(0.035)	(0.039)	(3.210)	(3.131)
γ[FM3-Γ-linear]	1.434	1.403	0.181	0.223	6.089	5.974
	(0.312)	(0.281)	(0.035)	(0.039)	(3.210)	(3.131)
R^2	0.994	0.995	0.999	0.999	0.901	0.908

Notes:
Standard Errors are in parentheses.
MPE is the Mean Prediction Error. The regression specification is given in equation (11).
$MAPE$ is the Mean Absolute Prediction Error. The regression specification is given in equation (12).
$ADMPE$ is the Absolute Deviation of Mean Prediction Error. The regression specification is given in equation (13).

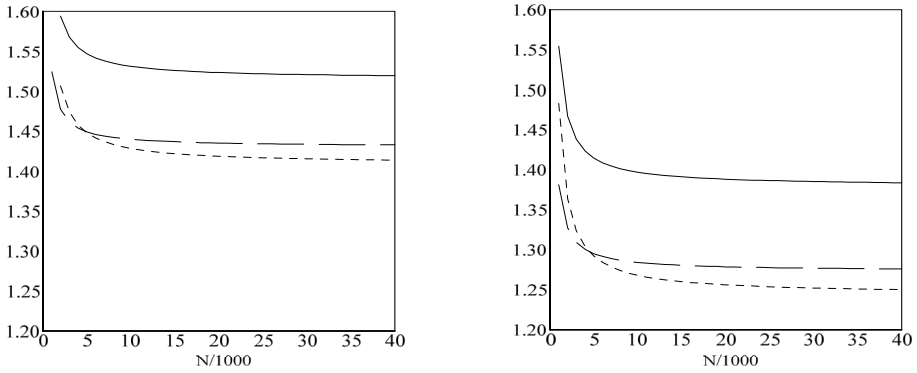
Figure 1: Properties of Prediction Errors

Untrimmed Prediction Sample | Trimmed Prediction Sample

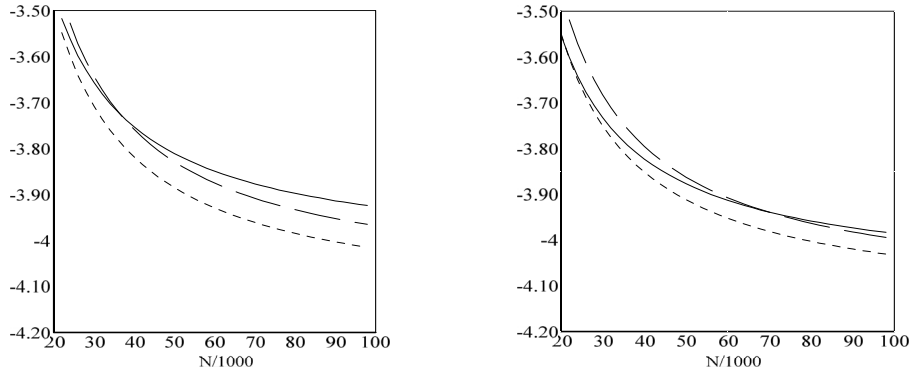
Mean Prediction Error



Mean Absolute Prediction Error



Absolute Deviations of Mean Prediction Error



Key: —— ols-linear — — Γ-linear - - FM2-Γ-linear