The modeling of medical expenditure data from a longitudinal survey using the Generalized Method of Moments (GMM) approach

Z. Hass[*], M. Levine[†], L.P. Sands[‡], J. Ting[§], and H. Xu[¶]

Abstract

Medical expenditure data analysis has recently become an important problem in biostatistics. These data typically have a number of features making their analysis rather difficult. Commonly, they are heavily right-skewed, contain a large percentage of zeros and often exhibit large numbers of missing observations due to death and/or the lack of follow-up. They are also commonly obtained from records that are linked to large longitudinal data surveys. In this manuscript, we suggest a novel approach to modeling these data through the use of GMM (Generalized Method of Moments) estimation procedure combined with appropriate weights that account for both dropout due to death and the probability of being sampled from among National Long Term Care Survey (NLTCS) subjects. This approach seems particularly appropriate due to the large number of subjects relative to the length of observation period (in months). We also use a simulation study to compare our proposed approach with and without the use of weights. The proposed model is applied to medical expenditure data obtained from the 2004-2005 NLTCS linked Medicare data base. The results suggest that the amount of medical expenditures incurred is strongly associated with higher number of activities of daily living (ADL) disabilities and self-reports of unmet need for help with ADL disabilities.

Keywords: GMM (Generalized Method of Moments), longitudinal data survey, IPW-GEE (Inverse Probability Weighting - Generalized Estimating Equations), modified sandwich estimator, medical expenditure data

1 Introduction

The rising medical expenditures have figured in the news rather prominently in the last several years and have prompted a substantial interest in the analysis of healthcare related data. Rising

[*]Purdue University

[†]Corresponding author. Department of Statistics, Purdue University 250 N. University St. West Lafayette, IN 47907-2066 tel. (765)496-7571 fax (765)494-0558 e-mail mlevins@purdue.edu

[‡]Virginia Tech

[§]Purdue University

[¶]IUPUI

health care expenditures for older adults have provoked a lot of concern recently. The goal of our research is to investigate whether unmet need for help with disabilities in activities of daily living (ADL) is associated with higher medical expenditures. If this is the case, such a connection would inform policy makers about resource planning for older adults with unmet ADL need. Unmet ADL need is a serious problem among older adults. 15% of older adults need help from others to complete basic activities of daily living (ADL)[1] such as bathing, dressing, eating, toileting, and getting around inside. However, nearly 20% of older adults who need ADL help report unmet need for assistance with their ADL [2, 3, 4]. Unmet need for ADL assistance is associated with increased healthcare utilization including hospitalization [5], re-hospitalization [6], and nursing home placement [7]. Unmet need for ADL assistance is also associated with increased risk for death [8] and, for many older adults, medical care expenditures increase significantly in the months before death [9].

The medical expenditure data are commonly available from a number of longitudinal surveys such as the NLTCS (National Long Term Care Survey). NLTCS is one of the longest running longitudinal surveys in the USA that has been ongoing for more than 30 years and it is linked to Medicare claims data. Since the detailed description of NLTCS has been given already in [10], we only offer a very brief one. Let the starting point of our observation period be Sept 1st, 2004. Beginning with this date, subjects begin entering the study starting from the day of the interview. The interview day can fall on any day between Sept 1st, 2004 and Dec 31st, 2005. As a result, we observe the straggled entry of subjects. Some subjects die during this follow-up period while others survive until Dec 31st, 2005 at which point the follow-up period stops for all subjects. Note that medical expenditures are taken from linked Medicare claims data; the reason follow-up was only conducted until the end of 2005 had to do with availability of the pre-linked Medicare records. The total medical expenditures are subdivided into a number of categories, e.g. durable medical equipment, hospital expenditures, skilled nursing facility, home health agency etc. For each subject, a number of covariates are available, such as unmet need for ADL disabilities, the number of ADL disabilities, and age. All of the covariates are binary. In addition to these two, some others were the diabetes status, the heart disease status etc. as well as demographic covariates such as age and gender. The total number of subjects was 2400 of which 467 died during the period of study.

The medical expenditure data tend to be rather complicated data and they present numerous statistical analysis challenges. First, they tend to be highly skewed to the right, with a relatively small proportion of patients incurring very high medical costs while the rest of patients hardly incurring any. Second, these data usually have a lot of missing observations either due to the lack of follow-up or death related dropout or both factors. Since the probability of death is related to expenditures due to significantly higher expenditures for many in the last months of life, the missing observations due to death cannot be viewed as MCAR (missing completely at random) which makes the analysis even harder. Third, the common simple random sampling (SRS) assumption cannot be used in their analysis since each observation point has the survey-related weight. We now discuss these issues in some extra detail.

The presence of skewness in the data, together with a large number of zeros, implies that

the choice of the modeling distribution may not be very straightforward. In general, such data cannot be viewed as generated by any particular continuous distribution. One of the approaches used to treat this problem is the use of OLS(ordinary least squares) regression with a positive shift at zero. For an overview of this approach, see, e.g. [11]. This approach has two significant shortcomings. First, the choice of the constant used to shift all of the observations away from zero tends to be rather arbitrary. Moreover, a retransformation back to the original scale is required in this case after the model has been fitted. Another possibility is to use the Tobit model which is, effectively, a censored normal regression that is based on the concept of latent variables. The genesis of this idea also goes back to econometric research; for details and a good overview, see [12] and [13]. The use of the Tobit model is problematic because it is very sensitive to violation of normality and heteroscedasticity assumptions (see [14] for more details). Moreover, Tobit model assumes that there is an underlying normal random variable that is censored due to some random mechanism; this implies, effectively, that zeros are not viewed as a valid response which is typically wrong in the medical cost data context.

Finally, the so-called two part model envisions a logit or probit GLM to model the probability of zero occurring while using another OLS or GLM to model the actual level of positive cost. This approach effectively models the fact that excessive zeros may be generated by a mechanism different from that of positive expenditures. Note that this amounts to the use of a degenerate mixture model where one of the components is concentrated at just one point. The two-part model has a long and distinguished history in various applications. A version of this model was used in 1970's by meteorologists for rainfall; see, e.g. [15], [16] and [17]. The first ever example of its use in economic context was probably [18]. Later, this model was widely used in health economics as a result of the well known Health Insurance Experiment conducted by RAND Corporation; in that context, the two-part model was introduced in [19] and [20]. [21] provided a good overall review of the widespread use of the two-part model for health care cost data. Note that in the cross-sectional context the two parts of the model may be fitted separately. An excellent recent work on the practical implementation of the two-part model in the cross-sectional context is [22]. This approach was later extended to the longitudinal data context; the first occurrence was, probably, in [23].

The presence of missing data because of death also creates a significant problem in the data analysis. Some options considered so far in the literature include using an estimated probability of survival, obtained using, for example, Kaplan-Meier approach, within a short subinterval as a weight and then summing up the mean total cost weighted by it over all of the intervals. Such an approach was first suggested in [24]. [25] managed to extend this approach to develop an estimator whose asymptotic distribution is independent of the choice of partition. A different approach models the hazard function of the terminal event (i.e. death) based on subject specific covariates as a part of the joint model. For more detailed discussion of this approach, see [26] and [10].

Finally, the fact that the data come from a longitudinal survey also needs to be taken into account. It has been long known that ignoring sampling weights can lead to severely biased parameter estimates with underestimated standard errors (see [27] for a detailed discussion of

this issue in the medical context). There has been relatively little research on how to account for sampling weights in biomedical modeling.

Our main goal in this manuscript is to estimate the extent of the influence that unmet ADL needs have on the amount of incurred expenditures within the follow-up interval. We also attempted to estimate the extent of this influence in the most unbiased way possible. The model is constructed to estimate net expenditures and also provide unbiased estimates of parameters (see the Table (4)), especially the one that reflects the influence that unmet ADL needs have on the probability of incurring expenditures as well as on the amount of expenditures. We constructed a model that allowed estimation of the influence of unmet ADL needs in the context of known contributors to older adults medical expenditures. These known contributors also include respondents baseline characteristics (e.g. age, ADL status). A number of subjects have died during the period of study and we incorporate the knowledge of their death in the form of Inverse Probability Weighting (IPW) procedure. Since observations missing due to death are clearly not Missing Completely At Random (MCAR), this information has to be incorporated to avoid possibly biased estimates of the cost [28]. Such a bias is typically a serious problem whenever the complete multivariate distribution of the data cannot be fully specified. It is possible to think of such a bias as resulting from two sources: one is the lack of information about the unmeasured disease severity between patients with different levels of ADL and the other is due to unbalanced nature of the data when some of the patients die during the period of study. Moreover, [29] (see p. 490) noted earlier that, ignoring the data that are not missing completely at random, such as MAR (Missing At Random), "...can potentially introduce bias in the estimates of regression parameters". Due to these two concerns, we are introducing inverse probability weighting into our model to account for the missingness pattern due to the dropout.

Our secondary goal in conducting this research is to propose a longitudinal model that can relate highly skewed medical expenditures data with substantial missing data proportion to unmet need for ADL and some additional covariates. The ultimate hope is that the conclusions obtained will be of use in public policy. The current manuscript is structured as follows. Section 2 introduces the population averaged model we use to describe total medical expenditures across all categories and a novel method we introduce in order to fit it. Section 3 is dedicated to illustrating how the model works with simulated data. Section 4 shows how the model performs with real data. Finally, Section 5 describes possible directions of future research.

2 Model

Suppose a group of n subjects is followed with the medical expenditures being observed on a monthly basis. For each subject i, $i = 1, \ldots, n$, we have m_i monthly expenditure recorded where $m_i \leq m = 13$ observations. The reason there may be less than m observations for some subjects is because some subjects die before the end of study. The total number of observations is $N = \sum_{i=1}^{n} m_i$. The observed expenditures of each subject are $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i,m_i})^{'}$ with Y_{ij} being expenditures in jth month for the ith subject. Also, denote $\mathbf{Y} = \left(\mathbf{Y}^{'}_1, \ldots, \mathbf{Y}^{'}_n\right)^{'}$. Of course, $Y_{ij} \geq 0$; moreover, for some of $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, $Y_{ij} = 0$; that happens when a subject doesn't incur expenditures in any of the categories. All the subjects that did not die during the period of study are not observed any longer after Dec 31st, 2005 which is a fixed date that does not depend on a subject and also does not depend on the medical costs. In other words, the experiment ends at a pre-specified time. Also note that the time a subject enters the study is not fixed to be Sept 1st, 2004 but can occur on any day after that date and until the end of the study period. Technically, the time of entry into the study for a subject is his/her initial interview day. However, we will consider the 1st day of the month that follows the interview month as the true time of entry since this is more in line with how the Medicare claims are processed. In such a setting, the medical expenditures data should be viewed as truncated on the left.

Some earlier work, e.g. [10], has only modeled the dependence of just one category - hospital expenditures - on a number of subject specific covariates. Our interest lies in modeling the total expenditures and, in particular, the influence of the number of ADL disabilities and unmet need for help with ADL disability on these expenditures. All of the covariates involved in this study were constant over time and so we denote $\mathbf{X}_i$ the vector of covariates for ith subject. We follow the approach similar to that used by [26] and [10] with several notable differences. First, we model jointly the probability of incurring total medical expenditures (that is, $P(Y > 0)$), and the amount of positive expenditures. Second, instead of the subject level approach used in [26] and [10], we use the population-averaged approach. In other words, instead of using subject-wise random effects to induce the necessary autocorrelation structure, the autocorrelation matrix for each subject is modeled directly. Also, unlike [26] and [10], we are using the GMM (Generalized Method of Moments) approach that is a generalization of the classical GEE (Generalized Estimating Equations) approach first pioneered in [30]. GMM was first proposed in econometric context in [31]. The classical GEE approach is very robust to misspecification of the subject-wise autocorrelation structure and can easily handle unbalanced designs. Also, the GEE choice is rather sensible when the number of subjects is large relative to the length of time period involved. However, GEE is known to be consistent only when the missing data are assumed to be MCAR which is clearly not the case for us. Moreover, our data are generated from what is effectively a mixture model that does not belong to an exponential family. The last two issues force us to consider a somewhat more general GMM which has been used extensively in econometrics for a long time.

Note also that our data have been obtained using a survey (NLTCS) with complicated

weigtht structure, and, moreover, the death of certain subjects produces the dropout effect. There are a number of ways to compensate for the dropout effect. One of them is to use only the so-called "complete cases" (subjects that didn't die until the end of the study period) with appropriate weights so that they account for "incomplete cases" as well. This approach is rather inefficient as it forces the researcher not to use a substantial proportion of the data. Therefore, we prefer to be able to weight the contribution of each subject at each month of observation to the total pseudo-likelihood of the model explicitly. This can be achieved using the Inverse Probability Weighting-GMM (IPW-GMM) approach; for detailed historical introduction see e.g. [29], Chapter 18. One of the few references in the literature using the GEE approach for data with excessive zeros is [32]; they consider just clustered (rather than specifically longitudinal) data that are not weighted in any way. [32] also uses an independent working correlation assumption and only takes into account the intrasubject correlation when computing "sandwich" estimators of standard errors of estimated parameters.

We assume that Y_{ij} may be equal to zero with a non-zero probability that depends on the covariate vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$. Let γ be the $p \times 1$ vector of parameters and denote $p(\mathbf{X}_i, \gamma)$ the probability of expenditures Y_{ij} being equal to zero. A number of choices are available to model this probability, e.g. the logistic or a probit model. To make our discussion simpler, we use a logistic model whereby

$$p(\mathbf{X}_i; \gamma) = \frac{\exp(\mathbf{X}_i'\gamma)}{1 + \exp(\mathbf{X}_i'\gamma)}$$

which implies that logit $P(Y_{ij} > 0|\mathbf{X}_i) = \mathbf{X}_i'\gamma$. Since the positive expenditures are highly skewed to the right, a right skewed distribution should be used to model it. A gamma family represents a convenient choice, including exponential and χ^2 distributions as special cases. We parameterize gamma density as $f(y; \lambda, \nu) = \frac{\lambda}{\Gamma(\nu)}(\lambda y)^{\nu-1}e^{-\lambda y}$ defined for $y \geq 0$. This parameterization (see, for example, [33]) is commonly used for purposes of generalized linear modeling. It is rather convenient since the mean is, then, $\frac{\nu}{\lambda} = \mu$ and the variance is $\frac{\nu}{\lambda^2} = \mu^2\frac{1}{\nu}$ with $\frac{1}{\nu}$ being the dispersion parameter. There is some empirical evidence from various applications that the gamma distribution with the constant dispersion parameter may be insufficiently heavy tailed for healthcare modeling purposes; thus, we assume that, in general, the dispersion parameter may depend on a set of covariates through the log link. We use the log link to model the mean expenditure of ith subject in jth month as well. For simplicity, we assume that the covariate vector used to model the mean and the dispersion parameter is the same as that in the logistic model above though this need not always be true. Thus, letting δ and ρ be respective parameter vectors and $\eta = \frac{1}{\nu}$ the dispersion parameter, we define $\log \mu_{ij} = \mathbf{X}_i'\delta$ and $\log \eta_{ij} = \mathbf{X}_i'\rho$.

For an observation Y_{ij}, denote $\xi_{ij} = 1$ if $Y_{ij} > 0$ and $\xi_{ij} = 0$ if $Y_{ij} = 0$. Denote $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$. Under the assumption of independence working model, we can write the full pseudo-likelihood of our model as

$$L(\mathbf{Y}|\gamma, \delta, \rho) = \prod_{i=1}^{n} L_i(\mathbf{Y}_i|\gamma, \delta, \rho) \tag{2.1}$$

where

$$L_i(\mathbf{Y}_i|\gamma,\delta,\rho) = \prod_{j=1}^{m_i}\{f(y_{ij}|\lambda_{ij},\eta_{ij})\}^{\xi_{ij}}\{\text{logit}^{-1}(\mathbf{X}_i'\gamma)\}^{\xi_{ij}}\{1-\text{logit}^{-1}(\mathbf{X}_i'\gamma)\}^{1-\xi_{ij}}$$

The above is only true if there are no missing observations and all observations come from a simple random sample. In practice, some subjects die before the end of the study period and this has to be reflected in a weight assigned to each observation. More specifically, we suggest treating each death as a terminal event and use the inverse probability weighting (IPW) procedure to take that into account. For ith subject, define $R_{ij} = 1$ if the ith subject has a recorded expenditure amount (whether 0 or positive) in jth month and 0 otherwise. Denote the vector of response indicators for ith subject $\mathbf{R}_i = (R_{i1},\ldots,R_{im})'$. Then, the occasion when the subject experiences the terminal event (death) is $D_i = 1 + \sum_{i=1}^{m} R_i$. For a complete case, where $m_i = m$ and the entire expenditure vector $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{im})'$ is observed, we have $D_i = m+1$. For an individual with an incomplete vector of $m_i < m$ responses, we only observe $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{im_i})'$ and $D_i = m_i + 1$.

We assume that the dropout due to death in jth month can be thought of as occuring "at random" in the classical sense - the probability of death only depends on the subject specific covariates as well as prior expenditures (observations) in months up to $j-1$st. In any month j, the probability of survival through the jth month for the ith subject is $p_{ij} = P(D_i > j|D_i \geq j) = P(R_{ij} = 1|R_{i1} = \cdots = R_{i,j-1} = 1)$. It is commonly assumed that $R_{i1} = 1$ for any subject i and, therefore, $p_{i1} = 1$. Due to the MAR (Missing At Random) assumption, it seems reasonable to assume that the probability of survival for ith subject in jth month depends on both subject-wise covariate vector $\mathbf{X}_i$ and on the expenditures in all of the prior months. Let $\mathbf{Y}_{ij} = (Y_{i1},\ldots,Y_{i,j-1})'$. Define the joint covariate vector $\mathbf{Z}_{ij} = (\mathbf{Y}_{ij}', \mathbf{X}_i')'$ - a vector that typically consists of all observed expenditures prior to jth month for ith subject as well as the subject specific covariates $\mathbf{X}_i$. The probability p_{ij} can then be modeled using, for example, logistic regression, as

$$\log\frac{p_{ij}}{1-p_{ij}} = \mathbf{Z}_{ij}'\alpha. \tag{2.2}$$

In order to compute the weight assigned to subject i in jth month, we take the inverse of *unconditional probability* of survival occurring in jth month which is $\prod_{k=1}^{j-1} p_{ik}$ and multiply it by the sample weight under the independence assumption. If we denote the sampling weight of the ith subject ω_i, the weights to be used are defined as $\omega_{ij} = \omega_i\left[\prod_{k=1}^{j-1} p_{ik}\right]^{-1}$.

Thus, if the ith individual had observed expenditures Y_{ij} in the jth month (whether zero or positive), that subject will receive weight ω_{ij} in that month. If, on the contrary, the subject has died before or in jth month, he will receive the weight of zero. For the ease of notation, let us introduce the joint parameter vector $\theta = (\gamma,\delta,\rho)$. Then, the log-pseudolikelihood will become

$$l(\mathbf{Y}|\theta) = \sum_{i=1}^{n} l_i(\mathbf{Y}_i|\theta)$$

where each

$$l_i(\mathbf{Y}_i|\theta) = \sum_{j=1}^{m_i} \omega_{ij} \left[\xi_{ij} [\log f(y_{ij}|\lambda_{ij}, \eta_{ij}) + \mathbf{X}_i^{'} \gamma] - \log(1 + e^{\mathbf{X}_i^{'} \gamma}) \right]$$

Note that, in our case, the marginal distribution of each observations Y_{ij} is a mixture that does not belong to an exponential family. Traditional GEE generally assumes that marginal distribution of observations Y_{ij} belongs to an exponential family; however, a slightly more general GMM (Generalized Method of Moments) (see, e.g., [31]) approach allows us to dispose of this assumption. GMM - based approach allows us, under weak assumptions on the true dependence within the sample, to obtain consistent and asymptotically normal estimators of the model parameters. As in the classical GEE method, this is done without taking into account the unknown dependence structure when formulating estimation equations.

Let the score function be $U(\mathbf{Y}|\theta) = \frac{\partial l(\mathbf{Y}|\theta)}{\partial \theta} = \sum_{i=1}^{n} U_i(\mathbf{Y}_i|\theta)$ where $U_i(\mathbf{Y}_i|\theta) = \frac{\partial}{\partial \theta} l_i(\mathbf{Y}_i|\theta)$. Then, the sequence of estimating equations is defined as

$$U(\mathbf{Y}|\theta) = 0$$

At this point, we need to invoke some results on convergence of GMM estimators in cases where the sample is not the random draw from the population of interest. For example, [34] showed that the asymptotic normality is still valid for a variety of situations where weights are used as propensity scores and are dependent both on covariates and on other observations (which corresponds to the missing at random case). By this result, the normalized difference $n^{-1/2}(\hat{\theta} - \theta)$ converges to a normal distribution with mean zero and variance Σ_θ that needs to be estimated. The usual sandwich estimator has to be modified due to the presence of estimated weights. The usual "sandwich" estimator that would be ordinarily used to estimate Σ_θ is

$$\hat{\Sigma}_\theta = \left\{ \frac{\partial}{\partial \theta} U(\mathbf{Y}|\hat{\theta}) \right\}^{-1} \left\{ \sum_{i=1}^{n} U_i(\mathbf{Y}_i|\hat{\theta}) U_i^{'}(\mathbf{Y}_i|\hat{\theta}) \right\} \left\{ \frac{\partial}{\partial \theta} U(\mathbf{Y}|\hat{\theta}) \right\}^{-1}$$

In order to estimate the weights p_{ij} (and, therefore, ω_{ij}), we need to run a logistic regression analysis of the "stacked" dataset that includes observations for all subjects. This involves solving the following system of equations:

$$S(\alpha) = \sum_{i=1}^{n} S_i(\alpha) = \sum_{i=1}^{n} \sum_{j=2}^{m} R_{i,j-1} Z_{ij} (R_{ij} - \pi_{ij}) = 0.$$

The system of equations above represents the sum of subject-wise score functions for the logistic regression model (2.2) used to compute propensity weights. Note that only when $R_{i,j-1} = 1$ the ith subject's contribution is not equal to zero. In other words, only when in a given month $j-1$ the ith subject had recorded (possibly zero) expenditures, this subject contributes a nonzero term to its score function $S_i(\alpha)$. One can also say here that, after the death of ith

subject he/she does not contribute anything to its score $S_i(\alpha)$. In order to adjust for estimated weights, we need to change the sandwich estimator. More specifically, we replace $U_i(\mathbf{Y}_i|\hat{\theta})$ in the middle part of the "sandwich" estimator with the residual of the multivariate regression of $U_i(\mathbf{Y}_i|\hat{\theta})$ on $S_i(\alpha)$. Such a residual is

$$U_i^*(\mathbf{Y}_i|\hat{\theta},\hat{\alpha}) = U_i(\mathbf{Y}_i|\hat{\theta}) - \\ - \left(\sum_{i=1}^{n} U_i(\mathbf{Y}_i|\hat{\theta})S_i'(\hat{\alpha})\right)\left(\sum_{i=1}^{n} U_i(\mathbf{Y}_i|\hat{\theta})S_i'(\hat{\alpha})\right)^{-1} S_i(\hat{\alpha})$$

Taking the above into account, the modified estimator of $\hat{\Sigma}$ becomes

$$\hat{\Sigma}_\theta^* = \left\{\frac{\partial}{\partial\theta}U(\mathbf{Y}|\hat{\theta})\right\}^{-1}\left\{\sum_{i=1}^{n} U_i^*(\mathbf{Y}_i|\hat{\theta},\hat{\alpha})U_i^{*'}(\mathbf{Y}_i|\hat{\theta},\hat{\alpha})\right\}\left\{\frac{\partial}{\partial\theta}U(\mathbf{Y}|\hat{\theta})\right\}^{-1}$$

This amounts to effectively using a residual from a multivariate regression of $U_i(\mathbf{Y}_i|\theta)$ on $S_i(\alpha)$ in order to reduce variability of estimated $\hat{\theta}$. For details of this approach, see e.g. [29], Chapter 18.

3 Simulation study

In order to assess model performance a simulation study was performed. Data characteristics of interest are sampling weights, positive right skewed response with a point mass on zero, correlated response over time, dropout due to death, and the staggered entry into the study due to differing interview dates.

In order to mimic the analysis of real data, 5 independent binary covariates X_{ik}, $k = 1, \ldots, 5$, were generated from a Bernoulli distribution with the probability $p = 0.5$. Out of many ways of generating multivariate distribution with gamma marginals, we select the so-called Clayton copula. In brief, for a random vector $\mathbf{X} = (X_1, \ldots, X_d)'$ with continuous marginal distributions $F_i(x) = P(X_i \leq x)$, applying a probability integral transform results in a random vector $\mathbf{U} = (U_1, \ldots, U_d)' = (F_1(X_1), \ldots, F_d(X_d))'$ with uniform marginals. Then, the joint cumulative distribution of $(U_1, \ldots, U_d)$ is said to be a copula. The Clayton copula belongs to a specific class of copulas commonly called Archimedean copulas that allow modeling of dependence for an arbitrary high number of dimensions using only one parameter. For more details and an introduction to copulas, see e.g. [35]. In our cases, expenditures for a finite population of $50,000$ individuals were generated from a 13 dimensional Clayton Copula with the parameter $\theta = 2$. In order to mimic expenditures Y_{ij}, we back transform individual univariate uniform random variables using the appropriate inverse gamma CDF and then multiply by the necessary scale factor. Let $\Gamma^{-1}(u;\alpha)$ be the value of inverse Gamma CDF at u with the shape parameter α. Now, the resulting "synthetic" expenditures are correlated with the correlation coefficient

$\rho \approx 0.5$ and are represented as

$$Y_{ij} = \frac{1}{\nu} \exp\left\{\delta_0 + \sum_{k=1}^{5} \delta_k X_{ik}\right\} * \Gamma^{-1}\left(u_{ij}; \frac{1}{\nu}\right)$$

where u_{ij} is the random variate from the copula corresponding to X_{ij}, Y_{ij} is the expenditure of ith subject in jth month, and ν is the dispersion parameter,

In order to introduce a point mass on zero, the probability of no expenditure was calculated for each month of each subject. An indicator of non-zero expenditure was generated from a Bernoulli distribution with probability p_{ij}^{exp} that was obtained as

$$p_{ij}^{exp} = \frac{1}{1 + \exp\left\{-\left(\gamma_0 + \sum_{k=1}^{5} \gamma_k X_{ik}\right)\right\}}.$$

We used the same set of covariates here as the one used to model the distribution of expenditures Y_{ij}. If the indicator was one, the expenditure was left as it was, if the indicator was zero, the expense was set to zero, to simulate months with no expenditures. The introduction of zeros reduced correlation to levels consistent with the observed data ($\rho \approx 0.3$).

In order to introduce staggered entry due to interview date, a random start month was drawn from a multinomial distribution with $\pi = (0.1810, 0.3101, 0.3142, 0.1634, 0.0294, 0.0011)$ for months 1 to 6 to mimic the sample exactly. Unconditional probabilities of survival are calculated based on a logistic model that depends on the same 5 binary covariates as before and the expenditure of the previous month. We assume the dependence on the expenditure of ith subject in the previous month in order to simulate the MAR (Missing At Random) compliant dropout times due to death. Unconditional probabilities of not dropping out in the jth month for the ith subject are computed as a product of the probability of not dropping out $j-1$st month multiplied by the probability of survival through jth month. It is assumed that the probability of survival in the first month is 1. Therefore, the probability that ith subject dies in jth month is given by the expression

$$p_{ij}^{death} = p_{i,j-1}^{death} * \frac{1}{1 + exp\left\{-(\alpha_0 + \sum_{k=1}^{5} \alpha_k X_{ik} + \alpha_6 Y_{i,j-1})\right\}}.$$

Beginning with the month following the start month, a binary death indicator was sampled from a Bernoulli distribution with $P = p_{ij}^{death}$. For each subject, expenditures prior to the start month and after the month of death (if any) were dropped as being unobserved and not occurring respectively. In order to create sampling weights, data was structured to have 100 strata with 10 clusters each containing 50 subjects. A two stage sampling design was implemented. The first stage consisted of selecting a simple random sample of two clusters per each stratum. The probability of selection for each cluster was, therefore, equal to 0.2. At the second stage, a sample of four subjects per cluster was chosen with probability proportional

to size so that we ended up with a sample of size 800. The size measure for ith subject was defined similarly to [10], namely

$$S_i = \frac{(0.25 + 0.5\sum_{k=1}^{5} X_{ik})(0.5 + \delta_i)}{1 + exp(-0.001\bar{y}_i)}$$

where δ_i is an indicator variable equal to one if the i^{th} individual died before the end of the study and zero otherwise; $\bar{y}_i$ is the average monthly cost of ith subject. Note that this is an informative sampling scheme where the subjects who died during the period of study are oversampled. Sampling weights were calculated as the inverse of probability of selection. This was repeated 1000 times to create 1000 samples with 800 subjects each.

For each sample, a logistic model was fitted to estimate death weights, and then the IPW-GEE model was fitted and the Robust Sandwich Variance Estimator was used for standard errors. Coefficient estimates, variability of those estimates, mean of the Robust standard errors, and coverage probability of true parameter values are presented in Table (1). Note that in 3.3% of simulations, estimation procedure did not converge and the respective results were not included in the Table (1). It is clear that our approach seems to work rather well with simulated data in this informative sampling scheme. Note, in particular, an excellent coverage probability for all of the confidence intervals involved.

Since we estimate monthly expenditures in this setting, the presence of the dropout related weights does not change the values of estimated parameters greatly. We conducted a separate simulation where the weights used were just sampling weights, that is, $\omega_{ij} = \omega_i$ for each subject i. The resulting Table is almost identical to the Table (1) and is omitted in the interest of brevity. What is different, however, is the numerical stability of the estimation procedure. Whereas before, as we mentioned earlier, the estimation procedure did not converge in only about 3.3% of all cases, when the dropout related weights are omitted, this number rises to approximately 15%. We believe this is yet another reason to include the dropout related weights in our estimation procedure.

For comparison purposes, we also fit the same model to the data with no use of weights; therefore, the presence of missing data due to death as well as sampling weights are both ignored. The results are given in the Table (2). Note that the fitted coefficients are now biased, with the particularly pronounced bias present in estimated intercept of the medical expenditure model and in the dispersion parameter of the gamma distribution. The informative sampling scheme we are using results in oversampling of subjects with higher average expenditures and those who died during the observation period. If this element of the sampling design is ignored, one would expect a larger intercept in the medical expenditure model as well as a larger estimated dispersion parameter. The latter would account for increased variability in medical expenditures that was not properly accounted for by the sampling design.

4 Analysis of the NLTCS data

Included in the analysis are the 2400 respondents of the 2004 NLTCS community survey of whom 467 died during the observation period. Each respondent was followed for up to 13 months after the interview. Observation stopped with either death or being censored at the end of 2005. On average, respondents were followed for 10.66 months with a standard deviation of 2.46 months. Monthly medical expenditures in the follow-up period were obtained from the Medicare claims data by accumulating recorded expenditures within the same month. 86.12% of respondents had a nonzero total medical expenditure during the follow-up period.

A descriptive summary of monthly expenditures and some weighted sample characteristics is given in the Table (3). Ordinarily, SAS PROC GENMOD is used for the GEE approach to estimation of longitudinal data-based models. However, PROC GENMOD requires an explicit distributional assumption out of the (fairly short) list; our data have been generated by a mixture of right-skewed continuous gamma distribution and the degenerate point-mass distribution. Due to that, we decided to use PROC NLMIXED instead since it can approximately solve the necessary non-linear estimating equation.

For comparison purposes, we fit gamma models with both constant and non-constant dispersion parameters. We examine goodness of fit of these distributions to the positive monthly healthcare expenditures. More specifically, 1000 data sets are simulated for each model; afterwards, the expected quantiles based on the model are computed and plotted against sample quantiles . These plots show that, in both cases, the fit is relatively good although the heteroscedastic version does offer certain improvement by providing a more heavy tailed distribution. We also tried alternative heavy tailed distributions, such as inverse gamma and Weibull. Respective quantile to quantile plots showed a much worse pattern and are not shown here in the interests of brevity.

The weights have been fitted using the same set of covariates as the main model with the addition of the prior months medical expenses. That set of covariates $\mathbf{Z}_{ij}$ consists of moderate to severe ADL disability (3-5 categories), unmet need for ADL disability, age (viewed as a continuous varibale), gender, race (white or nonwhite), and medical expenses in $j-1$st month.

The results of our analysis are presented in (4) and (5). The first two parts of these models are very similar. All of the parameters that are statistically significant in one of them are significant in the other and vice versa. For convenience, we will refer to p-values from (4). They suggest that both increased ADL and the unmet need for medical assistance are strongly associated with the higher amount of incurred expenditures ($p < 0.0001$ and $p < 0.0008$, respectively). The increased ADL is also strongly predictive of the higher probability of incurring expenditures ($p < 0.0072$); however, the unmet need is not strongly predictive of that same probability ($p < 0.2473$). Both gender and race also seem to be strongly associated with the higher probability of incurring medical expenditures; females seem to incur costs less often than males ($p < 0.0001$) and whites seem to incur expenditures more often than people of other races ($p < 0.0002$). Finally, the older age is strongly associated with larger medical expenditures($p < 0.0007$).

5 Discussion

Our work continues the recently observed trend of utilizing complex survey data in statistical healthcare research. A particularly beneficial feature of a survey such as NLTCS is its linkage to Medicaid/Medicare claims which provides an opportunity to study the relationship between demographic characteristics and medical expenditures. Modeling medical expenditures has long been known to be difficult due to the presence of a large number of zeroes as well as the highly skewed nature of the non-zero part of observations and commonly present missing observations. All of these features need to be taken into account when analyzing the data.

A distinctive feature of our analysis is the use of a marginal model that is estimated using the GEE/GMM approach with inverse probability weighting. This approach has been less popular in the literature than the competing mixed modeling approach; [32] is one of the few available examples but they only modeled general clustered (and not specifically longitudinal) data that didn't come from a survey and were not subject to a dropout phenomenon. We found the GEE/GMM based approach particularly applicable in our case since our dataset had a large number of participants relative to the length of observations available; this latter fact makes GEE/GMM asymptotic results practically applicable. Its other appealing properties is robustness to misspecification of the data covariance structure as well as the ease of adapting it to the presence of dropout in the data under the MAR (Missing At Random) Assumption.

It is also of interest to note that, due to the "working independence" assumption that we have, in fact, been using in our GEE/GMM analysis, we can utilize any of the subject-wise covariates $\mathbf{X}_i$, even though some of them may have missing values (see, e.g. the discussion in [29] on p. 529. Unlike in a classical IPW method that is only using the so-called "complete cases" (only subjects who didn't die until the end of the study period), the IPW-GEE/GMM estimator that we are using in this research, does not disregard any of the available data.

The choice of a particular distributional model for the data analysis in this situation is rather complicated. A large number of different distributions have been used in practice before, e.g. the generalized gamma distribution [10], Pareto distribution [36] and several others. We conjecture that a combination of excessive zeros and a right-skewed heavy-tailed distribution may also be modeled as a member of a general exponential dispersion family, e.g. one of the so-called Tweedie family distributions (see e.g. [37]). Such an approach would avoid the necessity of using a mixture model to describe the data and may result in a simplified estimation procedure. Future research is needed to investigate such a possibility.

6 Acknowledgements

The work of M. Levine has been partially supported by the NSF grant DMS-1208994. The work of Laura P. Sands has been partially supported by the NIH grant NIH 5R01AG034160.

References

[1] Hung WW, Ross JS, Boockvar KS, Siu AL. Association of chronic diseases and impairments with disability in older adults: a decade of change? *Medical Care* 2012; **50**(6):501.

[2] Desai MM, Lentzner HR, Weeks JD. Unmet need for personal assistance with activities of daily living among older adults. *The Gerontologist* 2001; **41**(1):82–88.

[3] LaPlante MP, Kaye HS, Kang T, Harrington C. Unmet need for personal assistance services: Estimating the shortfall in hours of help and adverse consequences. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 2004; **59**(2):S98–S108.

[4] Sands LP, Wang Y, McCabe GP, Jennings K, Eng C, Covinsky KE. Rates of acute care admissions for frail older people living with met versus unmet activity of daily living needs. *Journal of the American Geriatrics Society* 2006; **54**(2):339–344.

[5] Xu H, Covinsky KE, Stallard E, Thomas J, Sands LP. Insufficient help for activity of daily living disabilities and risk of all-cause hospitalization. *Journal of the American Geriatrics Society* 2012; **60**(5):927–933.

[6] DePalma G, Xu H, Covinsky KE, Craig BA, Stallard E, Thomas J, Sands LP. Hospital readmission among older adults who return home with unmet need for adl disability. *The Gerontologist* 2013; **53**(3):454–461.

[7] Sands LP, Xu H, Joseph Thomas III SP, Craig BA, Rosenman M, Doebbeling CC, Weiner M. Volume of home-and community-based services and time to nursing-home placement. *Medicare & Medicaid Research Review* 2012; **2**(3).

[8] He S, Craig BA, Xu H, Covinsky KE, Stallard E, Thomas J, Hass Z, Sands LP. Unmet need for adl assistance is associated with mortality among older adults with mild disability. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 2015; :glv028.

[9] Hoover DR, Crystal S, Kumar R, Sambamoorthi U, Cantor JC. Medical expenditures during the last year of life: findings from the 1992–1996 medicare current beneficiary survey. *Health Services Research* 2002; **37**(6):1625–1642.

[10] Xu H, Daggy J, Yu D, Craig BA, Sands L. Joint modeling of medical cost and survival in complex sample surveys. *Statistics in Medicine* 2013; **32**(9):1509–1523.

[11] Basu A, Manning WG. Issues for the next generation of health care cost analyses. *Medical care* 2009; **47**(7_Supplement_1):S109–S114.

[12] Amemiya T. Tobit models: a survey. *Journal of Econometrics* 1984; **24**(1):3–61.

[13] Tobin J. Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society* 1958; :24–36.

[14] Gregori D, Petrinco M, Bo S, Desideri A, Merletti F, Pagano E. Regression models for analyzing costs and their determinants in health care: an introductory review. *International Journal for Quality in Health Care* 2011; .

[15] Cole J, Sherriff J. Some single-and multi-site models of rainfall within discrete time increments. *Journal of Hydrology* 1972; **17**(1):97–113.

[16] Todorovic P, Woolhiser DA. A stochastic model of n-day precipitation. *Journal of Applied Meteorology* 1975; **14**(1):17–24.

[17] Katz RW. Precipitation as a chain-dependent process. *Journal of Applied Meteorology* 1977; **16**(7):671–676.

[18] Cragg JG. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society* 1971; :829–844.

[19] Manning WG, Morris CN, Newhouse JP, Orr LL, Duan N, Keeler E, Leibowitz A, Marquis K, Marquis M, Phelps C. A two-part model of the demand for medical care: preliminary results from the health insurance study. *Health, Economics, and Health Economics* 1981; :103–123.

[20] Duan N, Manning WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* 1983; **1**(2):115–126.

[21] Mihaylova B, Briggs A, O'Hagan A, Thompson SG. Review of statistical methods for analysing healthcare resources and costs. *Health economics* 2011; **20**(8):897–916.

[22] Belotti F, Deb P, Manning WG, Norton EC, *et al.*. twopm: Two-part models. *Stata Journal* 2015; **15**(1):3–20.

[23] Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* 2001; **96**(454):730–745.

[24] Lin D, Feuer E, Etzioni R, Wax Y. Estimating medical costs from incomplete follow-up data. *Biometrics* 1997; :419–434.

[25] Bang H, Tsiatis AA. Median regression with censored cost data. *Biometrics* 2002; **58**(3):643–649.

[26] Liu L. Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine* 2009; **28**(6):972–986.

[27] Kneipp SM, Yarandi HN. Complex sampling designs and statistical issues in secondary analysis. *Western Journal of Nursing Research* 2002; **24**(5):552–566.

[28] Huang Y. Cost analysis with censored data. *Medical care* 2009; **47**(7 Suppl 1):S115.

[29] Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*, vol. 998. John Wiley & Sons, 2012.

[30] Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; :121–130.

[31] Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1982; :1029–1054.

[32] Lu SE, Lin Y, Shih WCJ. Analyzing excessive no changes in clinical trials with clustered data. *Biometrics* 2004; **60**(1):257–267.

[33] McCullagh P, Nelder JA. Generalized linear models. 1989; .

[34] Nevo A. Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics* 2003; **21**(1).

[35] Nelsen RB. *An introduction to copulas*, vol. 139. Springer Science & Business Media, 1999.

[36] Mullahy J. Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. *Medical Care* 2009; **47**(7_Supplement_1):S104–S108.

[37] Jorgensen B. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)* 1987; :127–162.

Simulation Results					
Parameter	True Parameter Value	Mean Estimate	SE of Estimates	Mean of SE	Coverage Probability
α_0	1.83	1.8402	0.1823	0.0357	93.3%
α_1	-0.18	-0.2005	0.1124	0.1060	92.0%
α_2	-0.21	-0.2024	0.1099	0.1060	94.6%
α_3	2.35	2.3358	0.1194	0.1174	92.9%
α_4	-1.18	-1.1800	0.1269	0.1162	92.2%
α_5	-0.22	-0.2198	0.1085	0.1070	93.5%
β_0	7.30	7.2970	0.1416	0.1261	90.7%
β_1	0.27	0.2847	0.0847	0.0812	93.8%
β_2	0.34	0.3397	0.0839	0.0812	94.2%
β_3	-0.34	-0.3442	0.0914	0.0834	92.7%
β_4	0.00	0.0103	0.0806	0.0796	94.8%
β_5	0.19	0.1985	0.0858	0.0811	92.7%
ν	0.93	0.9511	0.0642	0.0597	91.2%

Table 1: Results of simulation when the weights were used

Simulation Results					
Parameter	True Parameter Value	Mean Estimate	SE of Estimates	Mean of SE	Coverage Probability
α_0	1.83	1.8657	0.11198	0.1158	93.1%
α_1	-0.18	-0.1941	0.0765	0.0769	95.2%
α_2	-0.21	-0.1855	0.0779	0.0781	94.4%
α_3	2.35	2.3150	0.0945	0.0905	92.4%
α_4	-1.18	-1.1649	0.0804	0.0829	94.6%
α_5	-0.22	-0.2177	0.0785	0.0768	94.3%
β_0	7.30	7.4174	0.0768	0.0770	65.6%
β_1	0.27	0.2754	0.0549	0.0526	93.7%
β_2	0.34	0.3313	0.0536	0.0537	95.1%
β_3	-0.34	-0.3447	0.0531	0.0514	94.8%
β_4	0.00	0.0269	0.0530	0.0518	90.6%
β_5	0.19	0.1904	0.0509	0.0520	95.1%
ν	0.93	1.0667	0.0451	0.0457	13.6%

Table 2: Results of simulation when the weights were not used

Weighted sample characteristics of NLTCS respondents					
		N	Percent with positive cost	Mean positive monthly cost	Percent of subjects died(%)
Age					
	Less than 75 years	549	85.8	1963.1	9.7
	75 years or above	2175	87.5	2088.2	13.3
Gender					
	Male	791	85.6	2311.6	17.3
	Female	1933	86.5	1860.9	10.2
Race					
	White	2387	84.4	2473.8	10.0
	Other	337	86.5	1922.3	12.8
Diabetes					
	No	2074	85.3	1871.1	12.0
	Yes	642	88.8	2271.1	13.2
Emphysema					
	No	2513	86.0	1941.8	11.8
	Yes	211	89.2	2510.8	19.4
ADL					
	1-3 Limitations	1417	84.4	1742.2	7.8
	4-5 Limitations	1307	88.2	2247.0	17.2
Unmet ADL need					
	No	2126	85.8	1859.3	11.5
	Yes	598	87.6	2430.8	15.6

Table 3: Summary characteristics of the data

Figure 1: Quantile-quantile plot for monthly expenditures to compare the fit to homoscedastic gamma

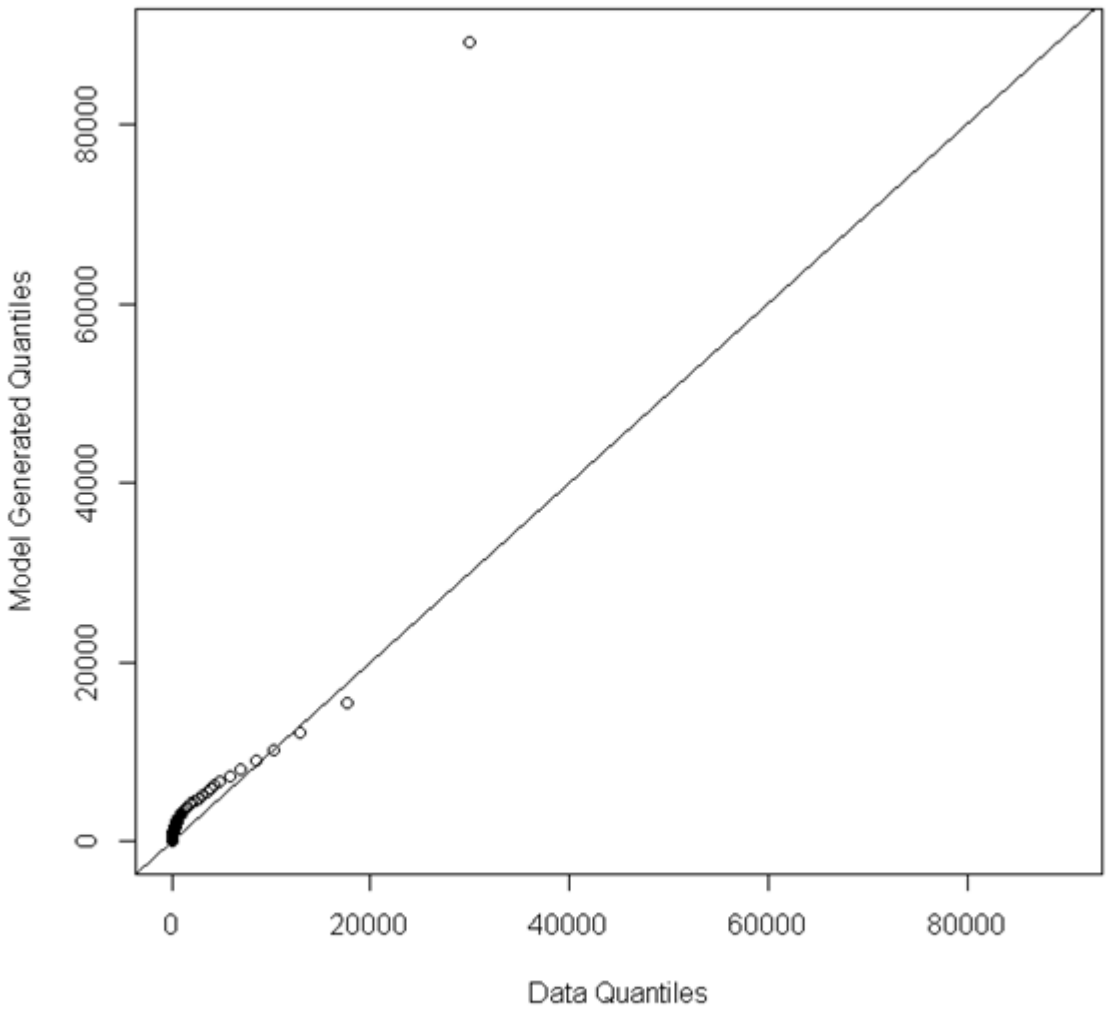
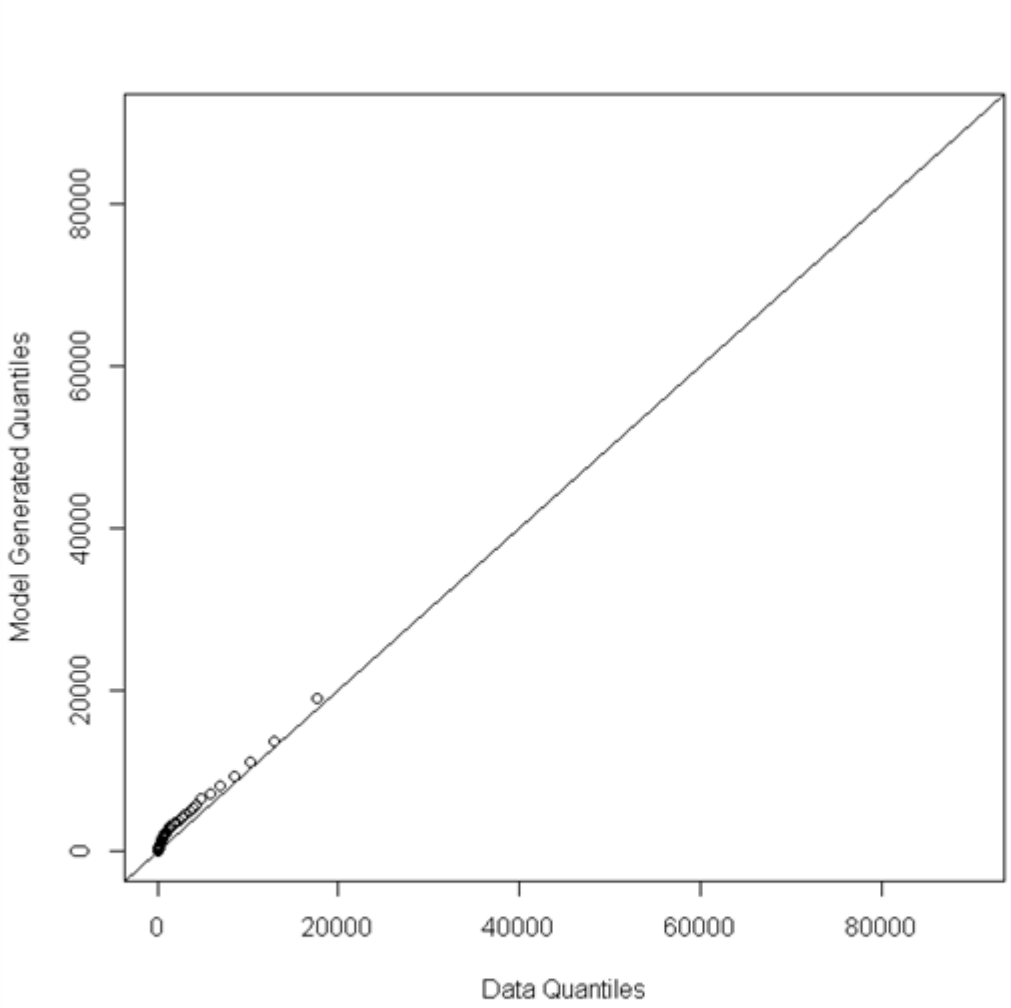


Figure 2: Quantile-quantile plot for monthly expenditures to compare the fit to heteroscedastic gamma



Parameter	GMM estimate	Standard error	p-value
Part I: incurring expenditures			
Intercept	8.8264	1.3426	0.0000
Unmet need	-0.1527	0.1320	0.2473
ADL 3-5	-0.3924	0.1460	0.0072
Age	-0.0848	0.0128	0.0000
Gender (female)	-1.7841	0.3900	0.0000
Race (white)	0.8102	0.1883	0.0002
Part II: amount of positive expenditures			
Intercept	6.6437	0.3196	0.0000
Unmet Need	0.2741	0.0817	0.0008
ADL 3-5	0.3825	0.0735	0.0000
Age	0.0140	0.0041	0.0007
Gender (female)	0.2058	0.0942	0.0290
Race (white)	-0.5118	0.1336	0.0001
Dispersion parameter	0.9103	0.1407	0.0000

Table 4: Parameter estimates for the population-averaged model of NLTCS data: homoscedastic gamma

Parameter	GMM estimate	Standard error	p-value
Part I: incurring expenditures			
Intercept	8.8267	1.3426	0.0000
Unmet need	-0.1528	0.1320	0.2472
ADL 3-5	-0.3924	0.1460	0.0072
Age	-0.0848	0.0128	0.0000
Gender (female)	-1.7840	0.3900	0.0000
Race (white)	0.8102	0.1883	0.0002
Part II: amount of positive expenditures			
Intercept	6.2887	0.3366	0.0000
Unmet Need	0.3140	0.0898	0.0005
ADL 3-5	0.4425	0.0677	0.0000
Age	0.0193	0.0044	0.0001
Gender (female)	0.2302	0.0974	0.0181
Race (white)	-0.5433	0.1318	0.0000
Part III: Dispersion parameter			
Intercept	2.0874	0.1970	0.0000
Unmet Need	-0.1166	0.0451	0.0097
ADL 3-5	-0.3217	0.0375	0.0000
Age	-0.0306	0.0024	0.0000
Gender (female)	-0.6493	0.0501	0.0000
Race (white)	0.2601	0.0535	0.0000

Table 5: Parameter estimates for the population-averaged model of NLTCS data: heteroscedastic gamma