MARGINALIZED TWO-PART MODELS FOR SEMICONTINUOUS DATA WITH APPLICATION TO MEDICAL COSTS

Valerie Anne Smith

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics.

Chapel Hill
2015

Approved by:

John Preisser

Brian Neelon

Amy Herring

Gary Koch

Matthew Maciejewski

ABSTRACT

Valerie Anne Smith: Marginalized Two-part Models for Semicontinuous Data with Application to Medical Costs
(Under the direction of John Preisser and Brian Neelon)

In health services research, it is common to encounter semicontinuous data characterized by a point mass at zero followed by a right-skewed continuous distribution with positive support. Examples include health expenditures, in which the zeros represent a subpopulation of patients who do not use health services, while the continuous distribution describes the level of expenditures among health services users. Semicontinuous data are typically analyzed using two-part mixture models that separately model the probability of health services use and the distribution of positive expenditures among users. However, because the second part conditions on a nonzero response, conventional two-part models do not provide a marginal interpretation of covariate effects on the overall population of health service users and non-users, even though this is often of greatest interest to investigators. Here, we propose a marginalized two-part model that yields more interpretable effect estimates in two-part models by parameterizing the model in terms of the marginal mean. This model maintains many of the important features of conventional two-part models, such as capturing zero-inflation and skewness, but allows investigators to examine covariate effects on the overall marginal mean, a target of primary interest in many applications. Using a simulation study, we examine properties of the maximum likelihood estimators from this model. We illustrate the approach by evaluating the effect of a behavioral weight loss intervention on health care expenditures in the Veterans Affairs (VA) health care system. We then extend this marginalized two-part model to clustered or longitudinal data structures by incorporating random effects. This longitudinal marginalized two-part model is fit following a fully Bayesian approach with non-informative or weakly informative prior distributions, and we illustrate it by analyzing the

effect of a copayment increase in the VA health system. Finally, using simulation studies, we compare the performance of the marginalized two-part model to commonly used one-part generalized linear models (GLMs) fit via quasi-likelihood estimation over a range of simulated data scenarios with varying percentages of zero-valued observations.

To my parents, Larry and Anne Smith,

and my grandparents, Hoyte and Myrtle Smith,

for being unconditionally supportive and always believing in me.

ACKNOWLEDGMENTS

I would like to thank my committee members, Drs. Amy Herring, Gary Koch, and Matthew Maciejewski, for their contributions and helpful comments. Many thanks to my advisors, Drs. Brian Neelon and John Preisser, for introducing me to marginalized models, for all the time spent discussing ideas, and for everything they taught me, both about modeling semicontinuous data and writing statistical papers. Thank you to Dr. Gary Koch, for providing invaluable advice, mentoring, and teaching over the past four years. I would also like to thank Drs. Maren Olsen and Matthew Maciejewski, whose support, encouragement, and teaching have been instrumental in developing my skills as a statistician and researcher.

Lastly, I would like to thank my family. To my husband, Ross McGurk, thank you for all of the time you spent listening to me talk about statistics and for all of your advice, encouragement, and understanding. To my parents, Larry and Anne Smith, thank you for your unwavering support and encouragement over the past 30 years. I would never be who I am nor where I am without all of the immeasurable and countless ways you have helped me to achieve my goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: LITERATURE REVIEW

1.1 Semicontinuous Data

In health services research, it is common to encounter semicontinuous data, such as medical expenditures (Manning et al. 1981; Duan et al. 1983), which are characterized by a point mass at zero followed by a right-skewed continuous distribution with positive support. In the case of medical expenditures, the point mass at zero represents a population of "non-users" who do not receive medical care in a given time interval and therefore have no medical expenditures; the continuous distribution, on the other hand, represents the level of expenditures among health services users given that expenditures were incurred. Considering the two defining components of such outcomes, semicontinuous data can be viewed as arising from two distinct stochastic processes: one governing the occurrence of zeros and the second determining the observed value conditional on it being a nonzero response. The first process is commonly referred to as the "occurrence" or "binary" part of the data, while the second is often termed the "intensity" or "continuous" part. Other examples of semicontinuous outcomes include hospital length of stay (Xie et al. 2004), health assessment scores (Su et al. 2009), and average daily alcohol consumption (Olsen and Schafer 2001; Liu et al. 2012).

The statistical modeling of such data provide unique challenges due to the "clumping" of observations at zero combined with the frequently right-skewed continuous distribution. A log transformation is often desired to normalize the distribution of the positive outcomes, but when employing such a transformation, one must decide how to address the zero values in the data. Some alternatives have been to either discard them or add a small constant to allow transformation. These solutions, however, are not ideal. Discarding zero values is not appropriate unless interest only lies in inference on the positive values, and adding a constant does not remove the discrete point mass, but rather only shifts it. Therefore, others

have chosen to model the probability of the outcome being zero separately from the value of the outcome conditional on it being positive in a "two-part" model. Each approach has advantages and disadvantages regarding statistical properties and interpretable results, and often one must compromise on one of these aspects in order to improve the properties of the other.

1.2 Two-Part Models for Independent Responses

1.2.1 Conventional Two-Part Model

There is extensive literature describing two-part mixture models for analyzing semicontinuous data. Aitchison (1955) initially highlighted the need for these two defining processes for unbiased estimation in applications involving estimation of expenditures and number of children per household. Deriving semicontinuous counterparts to many commonly used probability distributions, such as the exponential and log-normal, he defined these distributions as a mixture of the binary stochastic process and the continuous positive-valued process conditional on observing a positive response. In particular, for the log-normal distribution where the conditionally positive portion of the data follow the density

$$g(y) = \frac{1}{y\sigma\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2\sigma^2}\left[\ln(y) - \mu\right]^2\right\}, \quad y > 0,$$

he defined the mean and variance of the semicontinuous outcome as

$$\mathrm{E}(Y) = (1-\theta)\exp\left(\mu + \frac{1}{2\sigma^2}\right) \quad \text{and}$$

$$\mathrm{Var}(Y) = (1-\theta)\exp(2\mu+\sigma^2)\left[\exp(\sigma^2) - (1-\theta)\right],$$

where $\theta = \Pr(Y = 0)$.

Cragg (1971) extended this approach to the regression setting, modeling the binary and continuous components as functions of covariates. Manning and Duan (Manning et al. 1981;

Duan et al. 1983), as part of the RAND Health Insurance Experiment, introduced the most commonly used two-part model, termed throughout this document as the "conventional" two-part model. For data consisting of independent observations, the generic form of the conventional two-part model can be written as

$$f(y_i) = (1-\pi_i)^{1_{(y_i=0)}} \times [\pi_i g(y_i|y_i>0)]^{1_{(y_i>0)}}, \quad y_i \geq 0,\ i = 1,\ldots,n, \tag{1.1}$$

where $\pi_i = \Pr(Y_i > 0)$, $1_{(\cdot)}$ is the indicator function, and $g(y_i|y_i > 0)$ is any density function applicable to the positive values of Y_i, although the log-normal density is often chosen. This model is parameterized as

$$\eta(\pi_i) = \boldsymbol{z}_i'\boldsymbol{\alpha} \quad \text{and} \tag{1.2}$$

$$\mu_i = \mathrm{E}(\ln Y_i|Y_i > 0) = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{1.3}$$

where $\eta(\cdot)$ is an appropriate link function, typically a probit or logit function. When fitting this model to independent responses, the binary and conditionally continuous components of the likelihood are separable, and therefore, these two parts are fit separately. The binary component is often modeled using logistic regression, and the continuous component can be fit using standard regression models, such as the log-normal.

1.2.2 The Log-Skew-Normal Distribution

Because the log-normal distribution imposes a sometimes unrealistic condition of symmetry on the log-scale, alternative distributions such as the log-skew-normal have recently been proposed for the continuous part in an effort to relax these somewhat restrictive assumptions (Azzalini 1985; Chai and Bailey 2008). Azzalini (1985) first introduced the skew-normal distribution through the inclusion of a shape parameter, λ, that permitted skewness in a family of distributions related to the normal distribution. His family of skew-normal distributions offered strict inclusion of the normal density when $\lambda = 0$, was mathematically tractable, and

allowed for varying levels of skewness.

In the case of data that are positively-valued and right-skewed, such as medical expenditures, a log transformation can be useful to restrict the range of the predicted original values to the positive scale. Using this transformation, the log-skew-normal density of the positively valued observations becomes

$$g(y_i|y_i > 0) = \frac{2}{\omega y_i}\phi\left(\frac{\ln y_i - \xi_i}{\omega}\right)\Phi\left(\frac{\lambda}{\omega}(\ln y_i - \xi_i)\right).$$

with location parameter ξ_i, scale parameter $\omega > 0$, and shape parameter λ, all on the log scale, and where $\phi(\cdot)$ is the probability density function and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution.

Chai and Bailey (2008) showed the superior performance of the log-skew-normal (LSN) distribution compared to a log-normal distribution when conducting inference on the positive continuous part of positively skewed semicontinuous data. They additionally highlighted that the log-normal distribution is a special case of the LSN distribution, when $\lambda = 0$, so the appropriateness of the LSN vs. log-normal distribution can be assessed via a likelihood ratio test.

The generalized gamma distribution has also been proposed as a more general and flexible alternative to the standard log-normal distribution to model the continuous part of the data (Manning et al. 2005; Liu et al. 2010; 2012). This distribution takes as special cases the standard gamma, inverse gamma, Weibull, and log-normal distributions. Liu et al. (2012) compared the performance of models using the generalized gamma and LSN distributions for an analysis of alcohol-drinking outcomes from a clinical trial of a drug intended to reduce alcohol dependence. They found that the generalized gamma provided a better fit in this example, although different data sets could provide different preferred distributional assumptions. They also found the LSN distribution provided a better fit compared to a log-normal distribution.

1.2.3 Parameter Interpretation

Because they explicitly accommodate both data generating processes, two-part mixture models are an ideal choice for modeling semicontinuous data. Regardless of the distribution used, however, covariates in the second, or continuous, part of such two-part models are interpreted conditionally upon having observed a positive outcome. Consequently, attempts to combine these two parts to form the overall marginal mean effect of any covariate relies on specifying values for each of the other covariates in the model. As such, it is generally challenging to obtain a straightforward interpretation of covariate effects on the marginal mean in two-part models.

In many cases, however, investigators' main interest lies in examining such effects on the marginal mean in order to draw conclusions about the impact of predictors on the population as a whole. For example, in economic studies of system-wide health care expenditures, investigators and policy makers may wish to understand the average effect on medical expenditures of increasing specialty care copayments (Maciejewski et al. 2012a) or of bariatric surgery for weight loss (Maciejewski et al. 2010b; 2012b) on the entire affected or eligible populations rather than estimating separate effects for the probability of incurring expenditures and the level of expenditures given that any are incurred. In particular, an intervention may have one effect on the probability of occurrence but the opposite effect on the intensity given occurrence. In such cases, policy makers may be left without a true understanding of the overall population-level effect of such an intervention.

To achieve more interpretable effects, Mullahy (1998) and Buntin and Zaslavsky (2004) proposed using a one-part exponential conditional mean model to estimate effects of covariates on the marginal mean. While this one-part model provides interpretable estimates, it does not explicitly account for the zero-inflated nature of the data or provide investigators with estimates of covariate effects on the probability of occurrence. Thus, alternative models must be considered when interest lies in estimating both the binary component and the overall marginal mean.

1.3 Two-Part Models for Clustered Semicontinuous Data

Clustered semicontinuous data arise from many situations. In the example of medical expenditures, analysts may be interested in trajectories of a population's expenditures on health care over several years, or alternatively, may be interested in expenditures on prescription drugs incurred by patients clustered within physicians (Zhang et al. 2006). All of the issues related to model estimation of cross-sectional semicontinuous data are relevant to clustered semicontinuous outcomes as well. However, there are additional complications and considerations with clustered outcomes. As with any clustered outcome, the model estimation approach needs to incorporate the correlation of repeated measurements in addition to accounting for missing data due to loss of follow-up or death. Furthermore, in the case of longitudinal data, the distribution of longitudinal outcomes, and in particular, the proportion of zeros, is dependent upon the length of time interval under consideration. In many situations, particularly with health expenditures, the longer the time interval, the smaller the proportion of observed zeros in the expenditure distribution.

1.3.1 Conventional Two-Part Model for Clustered Semicontinuous Data

Olsen and Schafer (2001) first extended two-part models to longitudinal data. They proposed a logistic regression model with random effects for the binary part of the data combined with a linear mixed effects model for the log of the conditionally positive part and assumed that the random effects from these two models were jointly normally distributed and possibly correlated. This allowed the probability of occurrence at one time point to be associated with the level of the outcome given occurrence at another time point. Such a model can take the notation

$$\text{logit}(\boldsymbol{\pi}_i) = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{c}_i \quad \text{and} \tag{1.4}$$

$$\boldsymbol{\mu}_i = \text{E}(\ln \boldsymbol{Y}_i | \boldsymbol{Y}_i > 0) = \boldsymbol{X}_i^*\boldsymbol{\gamma} + \boldsymbol{Z}_i^*\boldsymbol{d}_i. \tag{1.5}$$

with

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{c}_i \\ \boldsymbol{d}_i \end{pmatrix} \sim N\left(\mathbf{0}, \Psi = \begin{pmatrix} \Psi_{cc} & \Psi_{cd} \\ \Psi_{dc} & \Psi_{dd} \end{pmatrix}\right)$$

Olsen and Schafer suggested a Laplace approximation with a Fisher scoring algorithm to obtain maximum likelihood estimates, and illustrated their model using data from a longitudinal study of middle and high school students on alcohol use. They additionally conducted a simulation study showing low bias and good coverage probabilities for the parameter estimates of both the fixed effects and variance components.

Tooze et al. (2002) proposed a very similar two-part model with correlated random effects, utilizing quasi-Newton optimization of the likelihood approximated by Gaussian quadrature rather than a Laplace approximation. This provided the ability to fit the model in standard statistical software packages, such as SAS (SAS Institute, Cary, NC), and they provided a SAS macro that calls procedures GENMOD and NLMIXED to fit such models. Rather than illustrating their model on a longitudinal data set as in Olsen and Schafer, they provided an example of its use for cross-sectional medical expenditure data that was clustered by household. Their method, however, could also be applied to longitudinal repeated measurements data.

The models proposed by Olsen and Schafer and Tooze et al. both account for potential correlation between the binary and continuous parts of the data. If, however, the occurrence and intensity are uncorrelated, the likelihoods are separable and thus each can be fit separately by maximum likelihood methods. This independence assumption can be very attractive as the inclusion of correlated random effects can introduce severe computational difficulties, at times so extreme that it may not even be possible to fit such a model. However, in many situations, it is quite reasonable to believe that the probability of occurrence, such as having any medical expenditures, at one time point may be related to the intensity, such as level of expenditures, at another time point. For example, patients who are more likely to incur expenditures may also have higher expenditures when they are incurred.

Su et al. (2009) discussed the bias introduced by incorrectly assuming independence be-

tween the binary and continuous parts of the model. Because the second part of a conventional two-part model includes only those who had a positive outcome, bias from informative cluster sizes arises as parameters in the binary part of the model influence the cluster size in the continuous part of the model. For example, if those who are more likely to have a positive outcome are also more likely to have a higher level of the outcome given occurrence, then higher levels of the outcome will be over sampled in the continuous model. Su et al. showed that an incorrect assumption of independence between the occurrence and intensity models can produce bias in the estimation of both the regression coefficients and variance components in the continuous part of the model. Further, the direction and size of this bias for most of the estimates relies on true values of the other parameters, including variance components, in both parts of the model. As such, it can be difficult to quantify in a general pattern.

Since the introduction of the correlated, longitudinal two-part model, others have extended it to additional situations. For example, Liu et al. (2008a) incorporated four parts rather than two parts, modeling separately the probability of incurring inpatient and outpatient expenditures and the level of each given they were incurred. In a different manuscript, Liu et al. (2008b) also extended two-part models to multi-level models, incorporating a third level of clustering and correlated random effects. The correlated, longitudinal model thus provides a strong foundation from which many more flexible models can be adapted, although complicated model structures can at times be hampered by computational challenges.

1.3.2 Extension to Bayesian Modeling

Fitting correlated two-part models in a maximum likelihood framework requires optimizing over often intractable, multidimensional integrals which can lead to severe computational difficulties. Because of this, Bayesian approaches have been considered for fitting multi-level models to semicontinuous data. Bayesian methods also offer the advantage of incorporating prior information when it is available and eliminate the need to rely on asymptotic properties. While the potential for using Bayesian methodology was briefly mentioned in prior literature, Zhang et al. (2006) first thoroughly developed and implemented a Bayesian approach for a

two-part hierarchical model. They fit nearly the same underlying model as Olsen and Schafer, but replaced the logit link function in the binary part with a probit link function for computational simplicity. Using non-informative prior distributions and Markov Chain Monte Carlo (MCMC) sampling, they used this two-part model to examine physician- and patient-level patterns in pharmaceutical expenditures among patients clustered within physician.

Similarly, Cooper et al. (2007) applied two-part models to longitudinal data, using Gibbs sampling MCMC methods and 'vague' prior distributions to analyze health care costs over time among individuals with early inflammatory polyarthritis. They compared the results of four models, including both one-part and two-part models, with varying specifications of random effects and distributional assumptions.

Bayesian approaches have been proposed for other extensions to two-part models as well. Ghosh and Albert (2009) developed a two-part model using penalized splines to model the effect of time and the time by treatment interaction. They fit their model using Gibbs sampling MCMC methods and illustrated it by analyzing clinical trial data on acupuncture for treating chemotherapy-induced vomiting in breast cancer patients. Additionally, Neelon et al. (2011) developed a Bayesian two-part growth mixture model to characterize the effect of increased mental health and substance abuse benefits in the Federal Employee Health Benefits Program on mental health use and expenditures. They used an MCMC algorithm to fit a two-part latent class model under weakly informative prior distributions for all parameters, and provided a simulation study showing low bias for all parameter estimates and good coverage rates for the 95% credible intervals. In short, Bayesian approaches to fitting a wide array of two-part models have been shown to maintain good statistical properties while providing computational simplicity and flexibility.

1.3.3 Two-Part Population Average Models for Clustered Data

It is well known that, in the presence of random effects, parameters estimates in generalized linear models have a subject-specific interpretation as opposed to a population average interpretation. Unless linear models with an identity link are used, the subject-specific param-

eter estimates differ in magnitude from their population average counterparts (Diggle et al. 2002). Most work in two-part model marginalization has been with regard to marginalizing over the random effects, converting subject-specific parameter estimates into population average estimates. Hall and Zhang (2004) proposed one such method for obtaining population average parameter estimates from zero-inflated models, utilizing an expectation solution (ES) algorithm, a generalization of the EM algorithm. This algorithm used generalized estimating equations (GEEs) in the S-step to estimate population average covariate effects while accounting for correlation within clusters. They applied their algorithm to several zero-inflated models, including zero-inflated Poisson, zero-inflated negative binomial, and zero-inflated censored log-normal models, in which they assumed that some zeros were true zeros and some were small positive values that were censored. However, due to the complexity of their algorithm, such estimation is not available in standard statistical software and therefore has not been widely implemented in practice.

Su et al. (2011) proposed a likelihood-based population average model for longitudinal semicontinuous data. Assuming a bridge distribution for the binary random intercept, as opposed to the ordinary normal distribution assumption, they provided a simple formula for converting the subject-specific binary parameter estimates into population average estimates based on an estimated parameter of the bridge distribution. Although later corrected, they incorrectly assumed that the continuous model would provide population average parameter estimates on the log scale due to using a linear mixed model with an identity link. In a correction, however, Tom et al. (2013) showed that, when correlation exists between the binary and continuous parts of the model, the population average parameter estimate is no longer equivalent to the subject-specific estimate when using an identity link. While there is no closed form for the conversion between subject-specific and population average parameter estimates in such scenarios, they provided mathematical bounds for the difference and suggested numerical techniques to calculate the conversion. Tom et al. also suggested marginalizing over the two parts of the two-part model to obtain estimates of the overall marginal mean, $\mathrm{E}(Y_{ij})$, when the outcome variable was log-transformed. A closed form was again not available, but they provided bounds within which the overall marginal mean would

lie and again suggested numerical evaluation.

As with independent data, an investigator's main interest often lies in examining covariate effects on the marginal overall mean, $\mathrm{E}(Y_{ij})$, in order to draw relevant policy conclusions about the impact of predictors on the population as a whole after accounting for clustering. While Tom et al. preliminarily addressed methods to calculate the overall mean, $\mathrm{E}(Y_{ij})$, under a log-transformation, their method did not estimate covariate effects on the overall mean. None of the two-part models for semicontinuous data provided in the literature provide parameter estimates that allow easy and interpretable estimation of such effects.

In the zero-inflated count literature, however, Long et al. (2014) proposed a marginalized model for zero-inflated Poisson (ZIP) regression. They parameterized their model in a two-part formulation, with the first part modeling the probability of an observation being a zero observed in excess of what is expected from a Poisson distribution. In the second part, they parameterized the model in terms of the overall mean, combining excess zeros with the Poisson-generated data. Utilizing this parameterization within the ZIP likelihood framework, they accounted for the zero-inflated nature of the data while also providing estimates of covariate effects on the overall mean, marginalized over the excess zeros in the distribution. Through several simulation studies, they showed low bias and good coverage probabilities for the model parameters and showed that, particularly in the presence of highly skewed covariates, their model out-performed standard Poisson regression and ZIP regression models. Illustrating their method with data from an intervention designed to reduce the number of risky sexual behaviors, they obtained an incidence density ratio for the intervention effect that was easily interpretable as the effect on the overall population mean number of risky sexual encounters. Hereafter, in this proposal, the term "two-part model" refers to two-part models for semicontinuous data rather than count data.

1.4 Comparison of One-Part vs. Two-Part Models

Two-part models, particularly for clustered data, can be computationally challenging to estimate. Additionally, conventional two-part models do not provide easily interpretable

effects of covariates on the overall marginal mean of both users and non-users, a quantity often of primary interest to investigators. Rather, two-part models provide estimates of covariate effects on the probability of having a positive outcome and on the level of the outcome conditional upon it being positive. The conventional two-part model thus may not be ideal for an analyst wishing to estimate the effect of covariates on the overall marginal mean. One-part models, on the other hand, incorporate both the zero and positively continuous values as arising from the same stochastic process and permit interpretation of covariate effects on the overall mean. One-part models typically take one of two general forms. In one form, a small constant is added to the outcome to ensure all values are positive and the outcome is then transformed to minimize skewness. Most commonly, the log transform is used. Alternatively, a generalized linear model (GLM) can be utilized, often with a log link, to avoid transformation and the need to add a constant to all values. Because these models allow simpler computation and interpretation, it is therefore of interest to question whether one-part models may be possible to use for semicontinuous data without creating bias or sacrificing too much precision.

Duan et al. (1983), when introducing the conventional two-part model, compared the performance of multiple models in estimating medical expenses from the RAND health insurance experiment. With approximately 20% of the sample incurring zero expenses, they examined an analysis of variance (ANOVA) model on untransformed (i.e., original scale) medical expenses, analysis of covariance (ANOCOVA) model on untransformed expenses, a one-part ordinary least squares (OLS) model on the log of the medical expenses, adding \$5 to ensure all expenses were positive, a two-part model with a probit model for the probability of incurring expenses and an OLS model for the log of the expenses given that they were incurred, and a four-part model, which was similar to the two-part model but modeled inpatient expenses with a separate two-part model. Using statistical consistency and minimum mean squared error as the criteria for judging the performance of the models, they found the ANOVA and ANOCOVA models yielded highly imprecise and noisy results, while the one-part model produced inconsistent results. Their two-part model also produced inconsistent results for inpatient expenses, and they found their four-part model to be the most accurate

and precise. Using a split-sample analysis for cross validation and comparing mean squared forecast error among the models, they found that the two-part and four-part models were indistinguishable while performing significantly better than their one-part counterpart and the ANOVA and ANOCOVA models. Ultimately, they recommended the four-part model when needing to assess inpatient expenses in an analysis.

Diehr et al. (1999) examined the performance of one-part and two-part models using data from Washington State's Basic Health Plan, where 21% of the sample incurred zero expenses. They fit three one-part models, including OLS on raw-scale dollars, OLS on log-scale dollars plus \$1 to ensure positive values, and a generalized linear model (GLM) using a gamma distribution and log link. For the two-part models, they fit a logistic regression to estimate the probability of incurring expenses, then analogously to the one-part models, fit three models to the continuous, positive part of the data: OLS on raw-scale dollars, OLS on log-scale dollars, and a GLM using a gamma distribution and log link. Fitting these models to a randomly selected half of the data, they used the other half to obtain root mean squared error (RMSE) and mean absolute error (MAE) for the predicted individual expenses of the other half of the sample. They found that the one-part log-normal model did not perform well using either RMSE or MAE, but that the other models showed no noticeable difference. Because of the lack of differences found, they recommended using a one-part model when the goal of inference is to understand the effects of individual covariates on total overall cost or when it is to predict future costs. They recommended using the two-part model when one's goal is to understand the processes driving the decision to obtain care, and thus incur costs, and then the level of expenditures given that they are incurred. Madden et al. (2000) also examined one-part vs. two-part models using data from Washington State, estimating expenditures separately for public employees and their dependents, including 21% with zero expenses, and individuals enrolled in the joint federal/state Medicaid program, including 8.5% with zero expenses. They compared a one-part OLS model on raw dollars to a two-part model with a logistic regression combined with a GLM using a gamma distribution and log link. They found that the more complex two-part model clearly outperformed the untransformed OLS model.

Mullahy (1998) emphasized the need for analysts to consider their modeling approach when inferences on $\mathrm{E}(y|\boldsymbol{x})$ are of primary interest. In particular, he focused on two main issues that arise in such two-part models: removing the conditioning on $y > 0$ and re-transforming from $\ln y$ to y. Emphasizing that re-transformation, and thus inferences on $\mathrm{E}(y|\boldsymbol{x})$, can be greatly biased if the model error term is dependent on the covariates, he advocated consideration of a modified two-part model, specified as $\mathrm{E}(y|y > 0, \boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta}_M)$ or a one-part exponential conditional mean (ECM) model, specified as $\mathrm{E}(y|\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\zeta})$. Analyzing the number of doctor visits in a 12 month period among 36,111 individuals ages 25 to 64, which included 23.6% having zero visits, he examined the performance of the ECM model, fit via non-linear least squares, compared to the conventional and modified two-part models. He found that parameter estimates were quite similar among the models and conjectured that larger differences may be found in a sample with a larger proportion of zero values. Examining mean prediction error (MPE) and mean squared error (MSE), he found that the modified two-part model performed slightly better than the ECM model, and both performed better than the conventional two-part model.

Expanding on the ideas presented in Mullahy's manuscript, Buntin and Zaslavsky (2004) compared nine potential modeling strategies, including variations of both one- and two-part models, for estimating mean Medicare expenditures on a sample with 8.6% of individuals having zero expenditures. Among the one-part models examined were an untransformed (i.e., raw scale) OLS model and three GLMs fit with quasi-likelihood, using constant variance, variance proportional to the mean, and variance proportional to the square of the mean. For the two-part models, they fit the second part using OLS on the log-scale expenditures, using four different re-transformation methods, and a GLM fit with quasi-likelihood and an assumption of constant variance. All GLMs used a log link. To examine the performance of these models, they evaluated predicted expenditures among relevant subgroups compared to actual mean expenditures in these groups and mean squared error (MSE), mean absolute prediction error (MAPE), and mean squared forecast error (MSFE) using split sample cross-validation. They concluded that four of the proposed models fit well, including the two-part OLS model with two smearing factors, the one- and two-part GLMs with constant

variance, and the one-part GLM with variance proportional to the mean. Ultimately, they recommended that researchers not specifically interested in the probability of use begin by fitting one-part models citing that zero observations can be included in such models without difficulty. If the probability of use were of specific interest, or if the researcher were unable to find a suitably fitting one-part model, they then suggested proceeding to examine two-part models.

Cooper et al. (2007) extended this comparison to a longitudinal setting, comparing four models to predict the costs incurred over time by individuals with inflammatory polyarthritis. Using a Bayesian perspective with vague priors and Gibbs sampling MCMC methods, they compared two one-part log-normal models, one with a random intercept only and one with a random intercept and slope for year, and two two-part models. The one-part models used log-transformed expenditures as the outcome with $1 added to ensure positive values. The two-part models both used logistic regression with a random intercept for the first part, and the second parts were a log-normal model with random intercept and slope for year and a gamma regression with log link, also including a random intercept and slope. The models were fit to a random sample of 76% of the data, and the remaining 24% was used to assess the predictive abilities of the models. In the 76% learning sample, the percentage of individuals with zero expenditures ranged from 32% to 54% over the years of the study. The models were also compared using the Bayesian Deviance Information Criterion (DIC). Under both of these criteria, the two-part models compared favorably to the one-part models.

Given the mixed conclusions from prior literature, there is no clear solution as to under what scenarios one-part models may provide a suitable alternative to their two-part counterparts when one's goal is to make inferences regarding the effect of covariates on the overall marginal mean of semicontinuous data.

1.5 Proposed Marginalized Two-Part Model

The debate regarding one-part vs. two-part models highlights the conflicting demands associated with modeling semicontinuous data. On the one hand, modeling approaches must

appropriately account for the unique statistical properties of semicontinuous data, but on the other, investigators need model estimates that are interpretable for their policy questions of interest. Previously, methods have not existed that simultaneously accounted for the excess zeros and skewness while also providing easily interpretable estimates of covariate effects on the overall marginal mean, $\mathrm{E}(Y)$.

This dissertation develops a new marginalized two-part (MTP) model that overcomes many of the drawbacks of previous approaches, including difficulty in interpreting covariate effects on the overall mean, a target of primary interest in many studies. Rather than parameterizing the model in terms of the mean of the transformed, conditionally positive outcomes in the second part, the MTP model parameterizes covariate effects directly on the overall mean, $\mathrm{E}(Y)$, on the untransformed scale. This allows parameter estimates to be interpreted as the multiplicative effect on the overall mean rather than on the conditional mean of only the positive outcomes. Our approach also has the advantage of providing estimates of covariate effects on the probability of incurring a positive-valued outcome, as in the first part of two-part models, as well as accounting for the zero-inflated and skewed nature of many semicontinuous outcomes.

We extend the MTP model to longitudinal or clustered data via the inclusion of random effects. This model can be fit using maximum likelihood or Bayesian approaches, although we propose the latter to increase flexibility and overcome computational difficulties when modeling complex random effect structures. This approach provides easily computed predictions of the overall mean outcome, and the parameter interpretations obtained from the MTP model provide the same simple interpretation as those from the one-part GLMs without sacrificing statistical appropriateness. Thus, the MTP model can provide useful policy conclusions while remaining rooted in good statistical practice.

Finally, we conduct a simulation study to compare the performance of the MTP model to that of one-part GLMs fit with quasi-likelihood. GLMs rely less on parametric assumptions but fail to directly address the discrete point mass at zero, while the fully parametric MTP model requires stronger distributional assumptions but accounts for the clumping at zero.

With such trade-offs, it is natural to question under what conditions each modeling approach exhibits better performance. Assessing bias, test size, and coverage of nominal 95% confidence intervals for covariate effects and model predictions, we fit these models to data generated under varying distributions, proportion of zeros, and sample sizes to inform under what scenarios the models are appropriate and when they encounter difficulties.

The remainder of this document is divided into three chapters. Chapter 2 describes the MTP model for cross-sectional data, examines properties via a simulation, and applies the model to assess the effect of a behavioral weight loss program on health care expenditures among an obese population in the Veterans Affairs health care system. Chapter 3 extends the approach to the longitudinal setting by developing a MTP mixed model with correlated random effects to allow dependence between the probability of incurring a positive outcome and the level of the outcome. For inference, we adopt a Bayesian approach because it avoids the computational challenges imposed in frequentist estimation, such as Gaussian quadrature approximation. The Bayesian approach also has the advantage of incorporating prior information and avoiding reliance on asymptotic inference. We illustrate the longitudinal model using a study of the effect of a copayment increase in the Veterans Affairs health care system. Chapter 4 compares one-part models fit using quasi-likelihood with the MTP model under a variety of simulated data generating mechanisms to assess under which scenarios one may be able to fit the simpler one-part models without inducing excessive bias or sacrificing too much precision, or alternatively, when two-part models are needed for appropriate statistical inference. Chapter 5 provides a discussion and points to future areas of research.

CHAPTER 2: A MARGINALIZED TWO-PART MODEL FOR SEMICONTINUOUS DATA[1]

2.1 Introduction

In health services research, it is common to encounter semicontinuous data, such as medical expenditures (Manning et al. 1981; Duan et al. 1983), which are characterized by a point mass at zero followed by a right-skewed continuous distribution with positive support. In the case of medical expenditures, the point mass at zero represents a population of "non-users" who do not receive medical care in a given time interval and therefore have no medical expenditures; the continuous distribution, on the other hand, represents the level of expenditures among health services users given that expenditures were incurred. Considering the two defining components of such outcomes, semicontinuous data can be viewed as arising from two distinct stochastic processes: one governing the occurrence of zeros and the second determining the observed value conditional on it being a nonzero response. The first process is commonly referred to as the "occurrence" or "binary" part of the data, while the second is often termed the "intensity" or "continuous" part. Other examples of semicontinuous outcomes include hospital length of stay (Xie et al. 2004), health assessment scores (Su et al. 2009), and average daily alcohol consumption (Olsen and Schafer 2001; Liu et al. 2012).

There is extensive literature describing two-part mixture models for analyzing semicontinuous data. Aitchison (1955) initially highlighted the need for these two defining processes for unbiased estimation in applications involving estimation of expenditures and number of children per household. Deriving semicontinuous counterparts to many commonly used probability distributions, he defined these distributions as a mixture of the binary stochastic

[1]This chapter previously appeared as an article in Statistics in Medicine. The original citation is as follows: Smith VA, Preisser JS, Neelon B, et al. "A marginalized two-part model for semicontinuous data," *Statistics in Medicine*. December 2014; 33(28):4891-4903.

process and the continuous positive-valued process conditional on observing a positive response. Cragg (1971), Manning and Duan (Manning et al. 1981; Duan et al. 1983), and others extended this approach to the regression setting, modeling the binary and continuous components as functions of covariates. Most commonly, the binary part is modeled via logistic regression and the continuous component via a log-normal model. However, because the log-normal distribution imposes a sometimes unrealistic condition of symmetry on the log-scale, alternative distributions such as the log-skew-normal have recently been proposed for the continuous part in an effort to relax these somewhat restrictive assumptions (Azzalini 1985; Chai and Bailey 2008). More recent extensions include incorporating longitudinal data (Olsen and Schafer 2001; Tooze et al. 2002), assessing bias (Su et al. 2009), and examining alternative data transformations (Mullahy 1998).

Because they explicitly accommodate both data generating processes, two-part mixture models are an ideal choice for modeling semicontinuous data. When adjusting for covariates, these models typically include one set of parameters for the binary response and a second set for the continuous component conditional on a positive response. In particular, covariates in the second, or continuous, part are interpreted conditionally upon having observed a positive outcome. Consequently, attempts to combine these two parts to form the overall marginal mean effect of any covariate relies on specifying values for each of the other covariates in the model. As such, it is generally challenging to obtain a straightforward interpretation of covariate effects on the marginal mean in two-part models.

In many cases, however, investigators' main interest lies in examining such effects on the marginal mean in order to draw conclusions about the impact of predictors on the population as a whole. For example, in economic studies of system-wide health care expenditures, investigators and policy makers may wish to understand the average effect on medical expenditures of increasing specialty care copayments (Maciejewski et al. 2012a) or of bariatric surgery for weight loss (Maciejewski et al. 2010b; 2012b) on the entire affected or eligible populations rather than estimating separate effects for the probability of incurring expenditures and the level of expenditures given that any are incurred. In particular, an intervention may have one

effect on the probability of occurrence but the opposite effect on the intensity given occurrence. In such cases, policy makers may be left without a true understanding of the overall population-level effect of such an intervention.

To achieve more interpretable effects, Mullahy (1998) and Buntin and Zaslavsky (2004) propose using a one-part exponential conditional mean model to estimate effects of covariates on the marginal mean. While this one-part model provides interpretable estimates, it does not explicitly account for the zero-inflated nature of the data or provide investigators with estimates of covariate effects on the probability of occurrence. Thus, alternative models must be considered when interest lies in estimating both the binary component and the overall marginal mean.

We propose a new "marginalized" two-part model for semicontinuous data which yields more interpretable effect estimates in two-part models by parameterizing the model in terms of the marginal mean. This model maintains many of the important features of conventional two-part models, such as capturing zero-inflation and skewness, but allows investigators to examine covariate effects on the overall marginal mean, a target of primary interest in many applications. We also propose an extension to accommodate log-skew-normal data to relax the commonly used log-normal assumption for the continuous part of the model. We illustrate the approach by evaluating the effect of a behavioral weight loss intervention on health care expenditures in the Veterans Affairs (VA) health care system.

The remainder of the paper is organized as follows: Section 2.2 introduces the marginalized two-part log-normal model and extends it to a log-skew-normal distribution. Section 2.3 presents results from a simulation study highlighting important features of our method. Section 2.4 applies the approach to the behavioral weight loss program, and Section 2.5 provides a discussion and points to areas for future work.

2.2 Marginalized Two-Part Models for Semicontinuous Data

2.2.1 Conventional Two-Part Model

We begin with a review of the conventional two-part model presented in Cragg (1971), Manning and Duan (Manning et al. 1981; Duan et al. 1983) and elsewhere. For data consisting of independent observations, the generic form of the conventional two-part model can be written as

$$f(y_i) = (1-\pi_i)^{1_{(y_i=0)}} \times [\pi_i g(y_i|y_i>0)]^{1_{(y_i>0)}}, \quad y_i \geq 0,\ i=1,\ldots,n, \tag{2.1}$$

where $\pi_i = \Pr(Y_i > 0)$, $1_{(\cdot)}$ is the indicator function, and $g(y_i|y_i>0)$ is any density function applicable to the positive values of Y_i, although the log-normal density is often chosen. This model is parameterized as

$$\text{logit}(\pi_i) = \boldsymbol{z}_i'\boldsymbol{\alpha} \quad \text{and} \tag{2.2}$$

$$\mu_i = \text{E}(\ln Y_i|Y_i>0) = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{2.3}$$

When fitting this model to independent responses, the binary and conditionally continuous components of the likelihood are separable, and therefore, these two parts are fit separately. The binary component is often modeled using logistic regression, and the continuous component can be fit using standard regression models, such as the log-normal.

2.2.2 Marginalized Two-Part Model

To obtain interpretable covariate effects on the marginal mean, we propose the following *marginalized two-part model* that parameterizes the covariate effects directly in terms of the marginal mean, $\nu_i = \text{E}(Y_i)$, on the original (i.e., untransformed) data scale. The marginalized

two-part (MTP) model specifies the linear predictors

$$\text{logit}(\pi_i) = \boldsymbol{z}_i'\boldsymbol{\alpha} \quad \text{and} \tag{2.4}$$

$$\text{E}(Y_i) = \nu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}). \tag{2.5}$$

Parameter estimates can be obtained using standard optimization routines such as Newton-Raphson or Fisher scoring. Model-predicted means and standard errors can also be easily obtained under this parameterization in a single step by estimating $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ at the desired values of the covariates.

2.2.3 Comparison of Treatment Effect Estimates

Using the conventional model shown in equation (2.3), γ_j is interpreted as the effect of a unit increase in the jth covariate, x_{ij}, on the conditional mean of $\ln(Y_i)$ given Y_i is positive. In many applications, however, this interpretation has limited usefulness as it is only relevant for the population of health services users. Rather, interest often lies in estimating the effect of covariates $\boldsymbol{x}_i$ on the marginal mean of Y_i for the combined population of health services users and non-users; that is the effect of $\boldsymbol{x}_i$ on $\text{E}(Y_i)$ unconditionally. In the case of the log-normal distribution, that is the effect of $\boldsymbol{x}_i$ on

$$\text{E}(Y_i) = \nu_i = \pi_i \exp(\mu_i + \sigma^2/2) = \frac{\text{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}}{1 + \text{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}} \exp(\boldsymbol{x}_i'\boldsymbol{\gamma} + \sigma^2/2), \tag{2.6}$$

where σ^2 is the variance of Y_i on the log scale. Assuming $\boldsymbol{x}_i = \boldsymbol{z}_i$, as is commonly specified, it follows from (2.6) that the per-unit effect of the j-th covariate, x_{ij}, on the marginal mean is

$$\frac{\text{E}(Y_i|x_{ij} = j+1, \tilde{\boldsymbol{x}}_i)}{\text{E}(Y_i|x_{ij} = j, \tilde{\boldsymbol{x}}_i)} = \frac{1 + \exp\left[\tilde{\boldsymbol{x}}_i'\tilde{\boldsymbol{\alpha}} + \alpha_j \cdot j\right]}{1 + \exp\left[\tilde{\boldsymbol{x}}_i'\tilde{\boldsymbol{\alpha}} + \alpha_j \cdot (j+1)\right]} \exp(\alpha_j + \gamma_j), \tag{2.7}$$

where $\tilde{\boldsymbol{x}}_i$ is $\boldsymbol{x}_i$ with x_{ij} removed and $\tilde{\boldsymbol{\alpha}}$ is $\boldsymbol{\alpha}$ with α_j removed. Thus, unless $\alpha_j = 0$, one must

specify fixed values for the remaining covariates in order to obtain a marginal interpretation for the effect of x_{ij}. Further, to obtain confidence intervals or formal inference on this marginal effect, the delta method or resampling techniques must be employed.

Using the MTP model as parameterized in equation (2.5), $\boldsymbol{\beta}$ is estimated for the entire population as opposed to $\boldsymbol{\gamma}$, which is conditional on $Y_i > 0$. Unlike γ_j in the conventional model, $\exp(\beta_j)$ can be interpreted as the multiplicative effect on the unconditional marginal mean, ν_i, when covariate x_{ij} increases by one unit. In other words, the left-hand-side of (2.7) equals $\exp(\beta_j)$ under model (2.5). Unlike the conventional model, standard errors and confidence intervals for covariate effects on the marginal mean are easily obtained as part of the standard model output.

There are other important distinctions between the models. In particular, when the model includes ancillary covariates with no interactions, the MTP model assumes a homogeneous treatment effect on $\mathrm{E}(Y_i)$ whereas the conventional model yields heterogeneous effects that depend on the specific values of the additional covariates, potentially creating misleading results. As an illustrative example, we generated a simulated dataset of sample size 10,000 using the following specification:

$$\mathrm{logit}(\pi_i) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} \quad \text{and}$$

$$\mathrm{E}(Y_i) = \nu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}),$$

where $x_{i1} \sim N(50, 100)$ and $x_{i2} \sim \text{Bernoulli}(0.5)$. We specified parameters values as $\alpha_0 = 14.4$, $\alpha_1 = -0.3$, $\alpha_2 = 1.6$, $\beta_0 = 5$, $\beta_1 = 0.05$, and $\beta_2 = 1.1$ and assumed Y_i followed the standard two-part distribution given in equation (2.1) with a log-normal density specified for $g(y_i|y_i > 0)$. Under this scenario, the true multiplicative "treatment effect" on $\mathrm{E}(Y_i)$ of x_{i2} taking a value of 1 versus 0 is $\exp(\beta_2) = \exp(1.1) = 3.0$. Because the ratio $\mathrm{E}(Y_i|x_{i2} = 1, x_{i1})/\mathrm{E}(Y_i|x_{i2} = 0, x_{i1}) = \exp(\beta_2)$ does not rely on specification of other covariate values, note that this treatment effect is the same regardless of the values for x_{i1}. Next, using equation (2.7), we estimated treatment effects under the conventional model at the first, second, and

third quartiles of x_{i1}, taking values 43, 50, and 57, respectively. Under the conventional model, the estimated effects of an increase in x_{i2} on $\mathrm{E}(Y_i)$ were multiplicative increases of 2.5, 4.4, and 8.3, respectively, while the true multiplicative increase was 3.0 regardless of the value of x_{i1}. It is worth pointing out that when treatment effect heterogeneity truly exists, the MTP model can accommodate this through the systematic inclusion of interactions, which should be driven by subject-matter considerations. While the conventional model also accommodates the systematic inclusion of subject-matter driven interactions, it imposes an arbitrary heterogeneity that always exists unless one omits the treatment covariate from the binary part of the model.

Because the interpretation of the marginal treatment effect from the conventional model relies on specified values of each other covariate in the model and includes arbitrary heterogeneity, "standardization" (Hernan and Robins 2014) is often used to obtain marginal treatment effect estimates by averaging across the observed heterogeneous estimates. In this approach, after modeling is complete, expected outcomes are estimated for each individual in the sample by first assuming they were a member of the treatment group then assuming they were a member of the control group; the overall treatment effect is then estimated as the difference in means of these constructed treatment and control groups. This approach has several disadvantages, however. First, bootstrapping or other resampling techniques must be employed to obtain standard errors and confidence intervals for estimated treatment effects. Further, the effect estimates are averaged over the sample distribution of the observed values of the other covariates in the model. Thus, if the sample distribution of these covariates does not represent the distribution in the target population, the treatment effect estimate is sample-specific and not as easily generalized to the overall population. This is not likely to be a problem in large datasets where the sample covariate distributions accurately represent those in the overall population. However, in smaller samples, the covariate distributions may not be representative of the population as a whole, leading to biased inferences. Additionally, estimates obtained via standardization lack generalizability to other populations which vary in the distribution of these covariates. As a result, the conventional two-part model lacks appeal when the objective is to estimate the effect of a covariate on the marginal mean. When

using the MTP model, however, estimation of the treatment effect does not require averaging over the observed values of the other covariates in the sample; therefore, these estimates provide much greater generalizability and ease of computation.

This is not to say that the MTP model should be preferred over the conventional model in all cases. Indeed, when the primary target of inference is $\mathrm{E}(Y_i|Y_i > 0)$, the MTP model engenders arbitrary heterogeneity and provides less interpretable estimates on the conditional mean of Y among the positive values. Ultimately, the choice between models should be guided by the aims of the analyst. If the aim is to model treatment effects on $\mathrm{E}(Y_i)$ in the presence of confounders, one should use the MTP model; on the other hand, if the target of inference is $\mathrm{E}(Y_i|Y_i > 0)$, the conventional model should be used.

2.2.4 Marginalized Two-Part Log-Normal Model

When modeling semicontinuous data, the continuous component is most frequently modeled using a log-normal distribution. The generic form of the two-part log-normal model for independent responses can be written as in (2.1) with $g(y_i|y_i > 0)$ taking the log-normal density function $\mathrm{LN}(\cdot; \mu, \sigma^2)$ with mean μ and variance σ^2 on the log scale. The marginal mean and variance of Y_i are then given by (Aitchison 1955):

$$\mathrm{E}(Y_i) = \nu_i = \pi_i \exp(\mu_i + \sigma^2/2) \quad \text{and} \tag{2.8}$$

$$\mathrm{Var}(Y_i) = \pi_i \exp(2\mu_i + \sigma^2) \left[\exp(\sigma^2) - \pi_i\right]. \tag{2.9}$$

The likelihood, parameterized in terms of π_i and μ_i, is:

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}|\boldsymbol{y}) = \prod_i (1-\pi_i)^{1_{(y_i=0)}} \left\{ \frac{\pi_i}{y_i\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(\ln y_i - \mu_i)^2 \right] \right\}^{1_{(y_i>0)}}.$$

In order to utilize this log-normal likelihood framework, the marginal mean in equation

(2.8) can be rearranged to solve for μ_i, yielding

$$\begin{aligned}\mu_i &= \ln \nu_i - \ln \pi_i - \sigma^2/2 \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} - \ln \pi_i - \sigma^2/2.\end{aligned}$$

Noting that

$$\pi_i = \frac{\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}}{1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}} \;\Rightarrow\; \ln \pi_i = \boldsymbol{z}_i'\boldsymbol{\alpha} - \ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}), \text{ and}$$

$$\ln(1-\pi_i) = -\ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}),$$

we can express the log-likelihood in terms of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and σ:

$$\begin{aligned}l(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma) = \sum_i -\ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}) + \sum_{y_i>0}\Bigg\{ &\boldsymbol{z}_i'\boldsymbol{\alpha} - \ln y_i - \frac{1}{2}\ln 2\pi - \ln\sigma \\ &-\frac{1}{2\sigma^2}\left[\ln y_i + \boldsymbol{z}_i'\boldsymbol{\alpha} - \ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}) + \sigma^2/2 - \boldsymbol{x}_i'\boldsymbol{\beta}\right]^2\Bigg\}\end{aligned}$$

with score equations

$$\boldsymbol{U}_i = \left[\begin{array}{ccc} \frac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma)}{\partial \boldsymbol{\alpha}} & \frac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma)}{\partial \boldsymbol{\beta}} & \frac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma)}{\partial \sigma} \end{array}\right]',$$

where

$$\begin{aligned}\frac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma)}{\partial \boldsymbol{\alpha}} &= \Bigg\{\frac{-\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}}{1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}} + \Bigg[1 - \frac{1}{\sigma^2}\left[\ln y_i + \boldsymbol{z}_i'\boldsymbol{\alpha} - \ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}) + \frac{1}{\sigma^2} - \boldsymbol{x}_i'\boldsymbol{\beta}\right] \\ &\qquad \cdot\left(\frac{1}{1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}}\right)\Bigg] 1_{(y_i>0)}\Bigg\}\boldsymbol{z}_i', \\ \frac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma)}{\partial \boldsymbol{\beta}} &= \left\{\frac{1}{\sigma^2}\left[\ln y_i + \boldsymbol{z}_i'\boldsymbol{\alpha} - \ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}) - \boldsymbol{x}_i'\boldsymbol{\beta}\right] + \frac{1}{2}\right\}\boldsymbol{x}_i', \\ \frac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma)}{\partial \sigma} &= \frac{-1}{\sigma}\Bigg\{1 - \frac{1}{\sigma^2}\left[\ln y_i + \boldsymbol{z}_i'\boldsymbol{\alpha} - \ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}) + \sigma^2/2 - \boldsymbol{x}_i'\boldsymbol{\beta}\right]^2 \\ &\qquad + \ln y_i + \boldsymbol{z}_i'\boldsymbol{\alpha} - \ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}) + \frac{\sigma^2}{2} - \boldsymbol{x}_i'\boldsymbol{\beta}\Bigg\}.\end{aligned}$$

With the conventional model the likelihood and score equations can be separated into two independent components: one for the binary part and one for the continuous part. In contrast, note that the score equations for the MTP model are not separable, and thus the binary and continuous parts are fit simultaneously. Model-based asymptotic standard errors are computed using Fisher's information matrix, $\mathfrak{I}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma)$ as

$$\text{s.e.}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\sigma}) = \sqrt{\text{diag}\left[\mathfrak{I}^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma)\right]}$$

with the maximum likelihood estimates substituted for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and σ.

2.2.5 Extension to the Log-Skew-Normal Distribution

While the log-normal distribution is suitable for many outcomes, it requires the somewhat restrictive assumption that the log-transformed outcome is symmetric and normally distributed, an assumption that is often violated in practice. We can relax this assumption by instead selecting for the positive responses a log-skew-normal distribution, which accommodates skewness through the inclusion of a shape parameter. Using the same linear predictors as in equations (2.4) and (2.5), the generic form of the two-part log-skew-normal model for independent data is given by:

$$f(y_i) = (1 - \pi_i)^{1_{(y_i=0)}} \times [\pi_i \text{LSN}(y_i; \xi_i, \omega, \kappa)]^{1_{(y_i>0)}}, \quad y_i \geq 0, \ i = 1, \ldots, n,$$

where $\text{LSN}(\cdot; \xi_i, \omega, \kappa)$ denotes the log-skew-normal (LSN) distribution with location parameter ξ_i, scale parameter $\omega > 0$, and shape parameter κ, all on the log scale, given by

$$g(y_i | y_i > 0) = \frac{2}{\omega y_i} \phi\left(\frac{\ln y_i - \xi_i}{\omega}\right) \Phi\left(\frac{\kappa}{\omega}(\ln y_i - \xi_i)\right).$$

The marginal mean and variance of Y_i for the MTP LSN model are given by:

$$\text{E}(Y_i) = \nu_i = 2\pi_i \exp\left(\xi_i + \frac{\omega^2}{2}\right) \Phi(\omega\delta) \quad \text{and}$$

$$\mathrm{Var}(Y_i) = 2\pi_i \exp\left(2\xi_i + \omega^2\right)\left[\exp(\omega^2)\Phi(2\omega\delta) - 2\pi_i\left(\Phi(\omega\delta)\right)^2\right],$$

where $\delta = \frac{\kappa}{\sqrt{1+\kappa^2}}$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal density. The likelihood, parameterized in terms of $\boldsymbol{\pi}$ and $\boldsymbol{\xi}$, is then

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\xi}, \omega, \kappa) = \prod_i (1-\pi_i)^{1_{(y_i=0)}} \left\{\frac{2\pi_i}{\omega y_i}\phi\left(\frac{\ln y_i - \xi_i}{\omega}\right)\Phi\left(\frac{\kappa}{\omega}(\ln y_i - \xi_i)\right)\right\}^{1_{(y_i>0)}},$$

where $\phi(\cdot)$ is the probability density function for the standard normal distribution. Thus, the log-likelihood in terms of $\boldsymbol{\pi}$ and $\boldsymbol{\xi}$ is:

$$\begin{aligned} l(\boldsymbol{\pi}, \boldsymbol{\xi}, \omega, \kappa) = \sum_{y_i=0} \ln(1-\pi_i) + \sum_{y_i>0}\Bigg\{&\ln\pi_i + \ln 2 - \ln(\omega) - \ln(y_i) + \ln\left[\phi\left(\frac{\ln y_i - \xi_i}{\omega}\right)\right] \\ &+ \ln\left[\Phi\left(\frac{\kappa}{\omega}(\ln y_i - \xi_i)\right)\right]\Bigg\}. \end{aligned} \tag{2.11}$$

In order to re-express the LSN likelihood as a function of $\boldsymbol{\beta}$, we first solve for ξ_i in terms of $\boldsymbol{\beta}$:

$$\begin{aligned} \xi_i =& \ln\nu_i - \ln 2 - \ln\pi_i - \ln\left[\Phi(\omega\delta)\right] - \frac{\omega^2}{2} \\ =& \boldsymbol{x}_i'\boldsymbol{\beta} - \ln 2 - \ln\pi_i - \ln\left[\Phi(\omega\delta)\right] - \frac{\omega^2}{2}. \end{aligned}$$

Plugging this into equation (2.11) above, the log-likelihood expressed in terms of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, ω, and κ is:

$$\begin{aligned} l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \kappa) = \sum_i -\ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}) + \sum_{y_i>0}\Bigg\{&\boldsymbol{z}_i'\boldsymbol{\alpha} + \ln 2 - \ln\omega - \ln(y_i) \\ &+ \ln\left[\phi\left(\frac{1}{\omega}\left(\ln y_i - \boldsymbol{x}_i'\boldsymbol{\beta} + \ln 2 + \boldsymbol{z}_i'\boldsymbol{\alpha} - \ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}) + \ln(\Phi(\omega\delta)) + \frac{\omega^2}{2}\right)\right)\right] \\ &+ \ln\left[\Phi\left(\frac{\kappa}{\omega}\left(\ln y_i - \boldsymbol{x}_i'\boldsymbol{\beta} + \ln 2 + \boldsymbol{z}_i'\boldsymbol{\alpha} - \ln(1+\mathrm{e}^{\boldsymbol{z}_i'\boldsymbol{\alpha}}) + \ln(\Phi(\omega\delta)) + \frac{\omega^2}{2}\right)\right)\right]\Bigg\}. \end{aligned}$$

Because the LSN reduces to the log-normal model when $\kappa = 0$, the choice between the log-normal model and the LSN model can be easily assessed using a likelihood ratio test. It should be noted that the LSN likelihood can be somewhat more sensitive to starting values with small sample sizes (say, n$<$ 30) when using software such as SAS PROC NLMIXED (SAS Institute, Cary, NC) that require initial values to be prespecified. This likelihood also has a stationary point at $\kappa = 0$ (Azzalini 1985); as such, 0 should not be provided as a starting value for κ in such estimation routines.

2.3 Simulation Study

To assess the performance of our proposed marginalized two-part model, we generated simulated data using the following specification:

$$\text{logit}(\pi_i) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} \quad \text{and}$$

$$\text{E}(Y_i) = \nu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}),$$

where $x_{i1} \sim N(50, 100)$ and $x_{i2} \sim \text{Bernoulli}(0.5)$. We specified parameters values as $\alpha_0 = 14.4$, $\alpha_1 = -0.3$, $\alpha_2 = 1.6$, $\beta_0 = 5$, $\beta_1 = 0.05$, and $\beta_2 = 1.1$. Using this specification, we generated 1,000 samples of size 10,000 under three scenarios with varying levels of skewness in the distribution:

i) $f(y_i) = (1-\pi_i)^{1(y_i=0)} \times \left[\pi_i \text{LN}(y_i; \mu_i, \sigma^2)\right]^{1(y_i>0)}$ with $\sigma^2 = 4$, or equivalently,
$f(y_i) = (1-\pi_i)^{1(y_i=0)} \times \left[\pi_i \text{LSN}(y_i; \xi_i, \omega, \kappa)\right]^{1(y_i>0)}$ with $\omega = 2$ and $\kappa = 0$;

ii) $f(y_i) = (1-\pi_i)^{1(y_i=0)} \times \left[\pi_i \text{LSN}(y_i; \xi_i, \omega, \kappa)\right]^{1(y_i>0)}$ with $\omega = 2$ and $\kappa = 2$; and

iii) $f(y_i) = (1-\pi_i)^{1(y_i=0)} \times \left[\pi_i \text{LSN}(y_i; \xi_i, \omega, \kappa)\right]^{1(y_i>0)}$ with $\omega = 2$ and $\kappa = 10$.

Under scenario (i), data were initially generated from a log-normal distribution with mean μ_i as shown in Section 2.2.4 and variance σ^2 on the log scale. Similarly, under scenarios (ii) and (iii), the data were generated from a LSN distribution with location parameter ξ_i as

defined in Section 2.2.5, scale parameter ω, and shape parameter κ, all on the log scale. Excess zeros were introduced in the Y_i's with probability π_i, resulting in 48% zeros. The log-normal data were generated using SAS 9.3 and the LSN data were generated using R version 2.15.2 (R Core Team 2012) using the `SN` package (Azzalini 2013). All models were fit using SAS 9.3 NLMIXED. The SAS code for fitting the log-normal and LSN MTP models is provided in Appendix A.

Table 2.1 shows the percent relative bias, median standard error, and coverage probability of each parameter from fitting each sample to both the log-normal and LSN MTP models. The log-normal marginalized model failed to converge 3 times when $\kappa = 0$ and once each when $\kappa = 2$ and $\kappa = 10$; the LSN model failed to converge 46 times when $\kappa = 0$, twice when $\kappa = 2$, and once when $\kappa = 10$. Likelihood ratio tests favored the LSN model in 3.2% of samples when $\kappa = 0$ and in 100% of samples when $\kappa = 2$ or $\kappa = 10$.

Under all scenarios, bias remained small and coverage probabilities were approximately 0.95 for all parameters except β_0. As skewness increased, bias increased for β_0 under the log-normal MTP model, and the coverage probability dropped to as low as 0.11 when $\kappa = 10$. While estimates of the remaining parameters would still be valid regardless of which model were used, when the data are skewed, the log-normal model is not appropriate for making predictions or estimating the overall mean, $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, due to the bias in β_0. Efficiency gains were also observed for the LSN MTP model when skewness was present; standard errors were somewhat smaller under the LSN model than the log-normal model when $\kappa = 2$ and even more so when $\kappa = 10$. These results indicate that the proposed MTP models provide unbiased estimates of regression coefficients. In the presence of skewness, however, the LSN model should be used to improve efficiency and yield unbiased predictions of the marginal mean, $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, since this is a function of β_0.

2.4 Analysis of MOVE! Intervention Data

The Veterans Affairs (VA) health care system implemented a system-wide weight loss intervention (MOVE!) beginning in 2006 to address the high prevalence of obesity among

VA patients (Kahwati et al. 2011). The high cost of obesity is well documented (Arterburn et al. 2005; Finkelstein et al. 2009), and the MOVE! intervention is the first behavioral weight loss program implemented across an entire health system. MOVE! was implemented as an unfunded mandate, so understanding its effect on average health care expenditures is important to guide program planning and refinement in VA. Assessment of the effect of MOVE! on the marginal mean of the entire VA population is also important for other health care systems that are also considering the adoption of behavioral interventions for reigning in the increasing costs of obesity. We use our marginalized two-part model to provide an estimate of the effect of the MOVE! intervention on the marginal mean expenditures among obese veterans.

The data for this analysis was drawn from a retrospective cohort study of obese VA patients eligible for MOVE! in fiscal years 2006-2009 who were identified from a longitudinal study of the VA cost of obesity. Data were obtained from the VA Corporate Data Warehouse (CDW) and the VA Outpatient Care File (OPC). As a part of a larger study, data were first obtained on all veterans who had received VA services and had a weight measurement in 2002 (N=3,365,004). This sample was then stratified into veterans who ever had one or more MOVE! clinic visits in 2006-2009 (MOVE! enrollees) or veterans who did not have a MOVE! clinic visit in this timeframe (non-enrollees).

Veterans were excluded from both cohorts if they were older than 70 in 2010, had a BMI of less than 30 kg/m^2 within 30 days of the index date, did not have sex data available, or had contraindications to MOVE! use during year of MOVE! initiation. Weight loss contraindications that warranted exclusion were central nervous system infections, organic brain syndromes or dementias, anorexia, anterior horn diseases, Huntington's disease, cirrhosis, dialysis, emphysema, neurological disorders, hepatitis, recent transplant surgery, or recent cancer treatment. Patients residing in nursing homes, hospice, or residential or adult day health care were also excluded.

To reduce the non-equivalence of MOVE! enrollees and non-enrollees due to imbalance in observed covariates, MOVE! enrollees and non-enrollees were first matched exactly on sex,

race (white or non-white), marital status (married or non-married), copay status (exempt vs. non-exempt), and veterans integrated service network (VISN) of residence. Then, potential matches were retained on the basis of BMI and comorbidity burden, assessed via the 2002 diagnostic cost group (DCG) score. Only matches with the same integer BMI measure occurring within 7 days of baseline measure of their respective MOVE! enrollee and the same DCG score (closest integer) were retained. The final cohort included 18,214 MOVE! enrollees and 18,214 non-enrollees.

The expenditure outcome of interest was total VA expenditures in the fiscal year following MOVE! initiation and was obtained from the VA Health Economics Resource Center. Expenditures for non-VA services were excluded as this analysis took a VA payer perspective. Total expenditures were inflation-adjusted to 2011 dollars using the general Consumer Price Index (CPI) because medical CPI does not adequately account for technological improvement, quality change and improved health outcomes (Berndt et al. 2002). The explanatory variable of primary interest was MOVE! initiation, which could occur any time between October 2005 and September 2009.

Descriptive statistics for the covariates and outcome are shown in Table 2.2. Of note, 17% percent of MOVE! enrollees and 14% percent of non-enrollees had zero health care expenditures in the year following initiation, yielding standard one-part log-normal or LSN models inappropriate. To assess the effect of MOVE! enrollment on health care expenditures in the following year, we fit the MTP model:

$$\text{logit}(\pi_i) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4}$$

$$\text{E}(Y_i) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})$$

where $x_{i1} = 1$ if individual i was enrolled in MOVE! and 0 otherwise, and we additionally adjusted for x_{i2}, x_{i3}, and x_{i4}, individual i's BMI, age, and DCG score, respectively. We fit this model twice, first assuming a log-normal distribution for the positive-valued observations, then assuming the more flexible LSN distribution. Table 2.3 presents the parameter estimates

and standard errors from the log-normal and LSN MTP models. Table 2.4 presents model-estimated mean expenditures from the LSN model at the quartile values of age, BMI, and DCG score for MOVE! enrollees and non-enrollees.

The estimates from the two models are quite similar, although a likelihood ratio test indicated that the LSN model was the more appropriate fit ($p < 0.0001$). Both models estimate an odds ratio of $\exp(-0.24) = 0.79$, indicating that the odds of incurring health care expenditures in the fiscal year following MOVE! enrollment were approximately 21% lower for those enrolled in MOVE! compared to non-enrollees with 95% confidence interval [CI] (17%, 26%). Despite the lower probability of incurring expenditures, however, we estimated from the log-normal MTP model that enrollment in MOVE! was associated with $\exp(0.1749) = 1.19$ times higher total health care expenditures on average in the following fiscal year with 95% Wald-type CI (1.16, 1.23). Similarly, the LSN MTP model estimated that MOVE! enrollment was associated with $\exp(0.1790) = 1.20$ times higher total health care expenditures on average in the following fiscal year with 95% CI (1.16, 1.23). While expenditures for non-enrollees remained lower than those of MOVE! enrollees, expenditures for both groups trended upward with increasing BMI and DCG score. Note that the estimated means for MOVE! enrollees at each quartile were 1.20 times higher than those of non-enrollees, reflecting the homogeneous model-estimated treatment effect across the distribution of age, BMI, and DCG score.

For comparison, we additionally fit the conventional two-part mixture model to these data. Using the same covariates as in the original analyses, we fit a logistic regression model to estimate the probability of incurring positive expenditures among the 36,428 individuals in our cohort and a log-skew-normal model on the subset of 30,847 individuals who had positive expenditures to estimate the level of expenditures conditional on occurrence. Thus, we fit the model:

$$\text{logit}(\pi_i) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4}$$

$$\text{E}(\ln Y_i | Y_i > 0) = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4}$$

Table 2.5 presents the regression estimates and standard errors from this model. The parameter estimates were similar to those estimated from the MTP model in Table 2.3. This reflects the fact that the percentage of zeros was not large, and hence the marginal mean was primarily driven by the positive expenditure values. Similar to our proposed model, the logistic regression suggested that MOVE! enrollment was associated with 21% lower odds of incurring health care costs in the following fiscal year compared to non-enrollees (95% CI [16%, 26%]). In contrast, the LSN model estimated that, conditional on incurring expenditures, those enrolled in MOVE! had 0.22 higher expenditures on the log scale on average than those not enrolled in MOVE! ($p < 0.0001$). However, with such conflicting results in the binary and continuous parts of this model, investigators are left without a clear sense of the combined overall effect of such an intervention on the average population cost. The MTP model, on the other hand, provides a single, easily interpreted estimate of the overall effect.

These MTP model results suggest that VA expenditures are not reduced in the year following MOVE! initiation, possibly because few veterans have sustained an intense participation in this behavioral weight loss program (Kahwati et al. 2011). This finding has important implications for VA policymakers needing to address the increasing incidence and prevalence of obesity among veterans. In particular, these results suggest that VA may need to introduce alternative weight management strategies to reduce expenditures among obese veterans or increase the effectiveness of MOVE! to induce expenditure reductions. It is possible that veterans' more recent (2012-2014) experience with MOVE! is more sustained and translates into VA expenditure reductions, but these findings suggest that enrollment in MOVE! in its initial four years was not associated with lower VA expenditures in the year following initiation, compared to non-enrollees.

2.5 Conclusion

We proposed a marginalized two-part model for semicontinuous data that allows investigators to obtain the population-average effect of covariates in the model. Our model directly parameterizes the covariates in terms of the population mean while still appropriately account-

ing for the excess number of zeros. While log-normal models are most commonly used, we also extended our MTP model to the broader and more flexible log-skew-normal distribution which allows asymmetry and skewness in the data and contains the log-normal distribution as a special case. In our simulation study, the maximum likelihood parameter estimates had near zero bias and good coverage properties in all scenarios except for the intercept from the log-normal model when fit to skewed data. As such, using either the log-normal or LSN MTP model should be reliable for estimating effects of covariates, although additional care should be used when predicting marginal means for specified covariate groups.

Using the LSN MTP model, we estimated that enrollment in the MOVE! weight loss intervention was associated with 20% higher average health care expenditures in the year after MOVE! initiation compared to a control group of non-enrollees. VA policymakers may need to refine the MOVE! program or introduce alternative behavioral weight programs to reduce expenditures among obese veterans. In contrast, the conventional two-part model found that MOVE! enrollment was associated with a decrease in the probability of incurring positive expenditures, but was also associated with an increase in the level of expenditures given they were incurred, leading to conflicting conclusions about the overall effect of the MOVE! program. Future directions for the MTP model could include extensions to clustered or spatially correlated data and applications to other fields, such as substance abuse or psychometric research. We are currently working to extend the MTP model to longitudinal data, allowing investigators to examine trends in the marginal mean over time. For example, it is possible that the effect of the MOVE! weight loss intervention may vary after additional years of enrollment, and estimating this effect would have important policy implications.

In short, the proposed MTP model provides a straightforward method for estimating covariate effects on the marginal mean of the population as a whole, which is not possible in conventional two-part models in a straightforward way. As such, it simplifies economic evaluations that are increasingly critical to understanding the return on investment of new interventions, policies and programs.

Table 2.1: Marginalized two-part model performance with 1,000 simulations and varying skewness

			Log-normal model			*Log-skew-normal model*		
κ	Parameter	True Value	% Relative Median Bias	Median Std Error	Coverage Probability	% Relative Median Bias	Median Std Error	Coverage Probability
0	α_0	14.4	0.09	0.2846	0.9488	0.08	0.2846	0.9486
	α_1	-0.3	-0.06	0.0058	0.9509	-0.05	0.0058	0.9497
	α_2	1.6	-0.08	0.0624	0.9468	-0.07	0.0624	0.9455
	β_0	5.0	0.11	0.1728	0.9478	0.44	0.1774	0.9486
	β_1	0.05	-0.41	0.0039	0.9549	-0.37	0.0039	0.9549
	β_2	1.1	-0.10	0.0585	0.9428	-0.12	0.0585	0.9444
2	α_0	14.4	0.12	0.2738	0.9489	0.13	0.2716	0.9429
	α_1	-0.3	-0.11	0.0055	0.9469	-0.14	0.0055	0.9419
	α_2	1.6	0.23	0.0604	0.9550	0.22	0.0599	0.9539
	β_0	5.0	-4.82	0.1218	0.4865	0.07	0.1222	0.9469
	β_1	0.05	-0.35	0.0028	0.9560	-0.19	0.0027	0.9539
	β_2	1.1	0.22	0.0419	0.9520	0.18	0.0400	0.9419
10	α_0	14.4	0.11	0.2672	0.9459	0.12	0.2298	0.9449
	α_1	-0.3	-0.12	0.0054	0.9429	-0.14	0.0046	0.9469
	α_2	1.6	0.19	0.0591	0.9520	0.05	0.0513	0.9570
	β_0	5.0	-6.81	0.1067	0.1111	0.0007	0.0764	0.9469
	β_1	0.05	-0.26	0.0025	0.9449	-0.06	0.0016	0.9489
	β_2	1.1	-0.08	0.0369	0.9580	0.09	0.0228	0.9309

Table 2.2: Means (SD) for MOVE! data

	MOVE! Enrollees (n=18,214)	Non-Enrollees (n=18,214)
Covariates		
Age	61 (9.3)	61 (9.3)
BMI	35.2 (3.9)	35.1 (3.9)
DCG Score	0.24 (0.17)	0.24 (0.17)
Outcomes		
% Positive Cost	83.2	86.2
Total Costs	7005 (18866)	6542 (18641)
Total Costs Among Users	8424 (20398)	7588 (19878)

Table 2.3: Marginalized two-part model results: MOVE! example

		Log-normal model		*Log-skew-normal model*	
	Parameter	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error
$\Pr(Y_i > 0)$					
Intercept	α_0	2.5653	0.1659	2.5680	0.1658
MOVE! Enrollment	α_1	-0.2381	0.0292	-0.2382	0.0292
BMI	α_2	-0.0321	0.0036	-0.0321	0.0036
Age	α_3	0.0080	0.0016	0.0080	0.0016
DCG Score	α_4	-0.3454	0.0849	-0.3457	0.0848
$\mathrm{E}(Y_i)$					
Intercept	β_0	9.1179	0.0875	9.1142	0.0872
MOVE! Enrollment	β_1	0.1749	0.0145	0.1790	0.0145
BMI	β_2	0.0079	0.0019	0.0084	0.0019
Age	β_3	-0.0176	0.0008	-0.0174	0.0008
DCG Score	β_4	1.2933	0.0446	1.2946	0.0444
	σ^2	1.4680	0.0118		
	ω			1.4123	0.0183
	κ			0.8426	0.0484

Table 2.4: LSN model-estimated means (standard errors) at quartiles of age, BMI, and DCG Score

	MOVE! Enrollees	Non-Enrollees
Age 55, BMI 32, DCG 0.11	$6303 (101)	$5270 (83)
Age 61, BMI 34, DCG 0.22	$6660 (90)	$5568 (73)
Age 66, BMI 37, DCG 0.32	$7127 (107)	$5958 (97)

Table 2.5: Conventional two-part LSN mixture model results: MOVE! example

	Parameter	Parameter Estimate	Standard Error
$\Pr(Y_i > 0)$			
Intercept	α_0	2.4906	0.1686
MOVE! Enrollment	α_1	-0.2369	0.0293
BMI	α_2	-0.0320	0.0036
Age	α_3	0.0093	0.0017
DCG Score	α_4	-0.3668	0.0859
$\mathrm{E}(\ln Y_i \mid Y_i > 0)$			
Intercept	γ_0	7.6503	0.0909
MOVE! Enrollment	γ_1	0.2153	0.0138
BMI	γ_2	0.0136	0.0018
Age	γ_3	-0.0186	0.0008
DCG Score	γ_4	1.3492	0.0421
	ω	1.4123	0.0183
	κ	0.8428	0.0484

CHAPTER 3: A MARGINALIZED TWO-PART MODEL FOR LONGITUDINAL SEMICONTINUOUS DATA

3.1 Introduction

In health services research, it is common to encounter semicontinuous data, characterized by a point mass at zero followed by a right-skewed continuous distribution with positive support. For example, medical expenditures typically include a point mass at zero representing a population of "non-users" who do not receive medical care in a given time interval and a continuous distribution that represents the level of expenditures among those who receive care. It is natural to view semicontinuous data as arising from two distinct stochastic processes. The first process, often referred to as the "occurrence" or "binary" part, governs the occurrence of zeros, while the second part, often referred to as the "intensity" or "continuous" part, determines the observed value conditional on it being nonzero.

There is an extensive body of work on analysis of cross-sectional semicontinuous data (Manning et al. 1981; Duan et al. 1983; Aitchison 1955; Mullahy 1998). Typically, semicontinuous data are analyzed using two-part models that explicitly accommodate both the binary and continuous data-generating processes. In the regression setting, the binary part is most commonly modeled via logistic regression and the continuous component via a log-normal model, although alternative distributions such as the log-skew-normal have been proposed for the continuous part to allow for additional flexibility (Azzalini 1985; Chai and Bailey 2008).

Two-part models have more recently been extended to accommodate longitudinal and clustered data (Olsen and Schafer 2001; Tooze et al. 2002). Typically, such models include a logistic regression with random effects for the binary part combined with a mixed effects model for the log-scale positive values. The random effects from these two parts are usually assumed to be jointly normally distributed and possibly correlated. Because the two pro-

cesses are allowed to be correlated, the two parts must be fit simultaneously rather than as two separate models. As a result, standard estimation approaches can encounter computational challenges when complex random effect structures are included in the model. Intensive procedures such as multi-dimensional quadrature approximations are needed, often creating long computational run times and convergence difficulties. Consequently, several authors have adopted a Bayesian inferential approach, which provides a flexible alternative ideally suited for more complex data structures (Zhang et al. 2006; Cooper et al. 2007).

Despite its widespread use, the conventional two-part model is limited in that it provides conditional interpretations for the regression coefficients in the continuous part of the model. Specifically, these parameters represent a location shift on the log scale that is conditional on both the random effects and on having observed a positive response. Therefore, it only refers to the population of "users", or those who incur a positive response, rather than the entire population of users and non-users, which is often of interest in health services studies. Further, the subpopulation of users changes over time in a longitudinal setting, so it is challenging to draw meaningful conclusions about the impact of covariates on expenditures for a fixed subpopulation of users. In many cases, however, investigators are primarily interested in examining the effect of covariates on the overall mean for the entire population in order to draw policy conclusions about their impact on that population and how that impact may change over time. For example, a medical system may wish to understand the long-term effect over many years of undergoing bariatric surgery on health care expenditures (Maciejewski et al. 2010b; 2012b) for the entire population eligible for surgery rather than estimating separate effects for the probability of incurring expenditures and the level of expenditures among those receiving care. Unfortunately, such marginal assessments are not easily obtained using the conventional two-part model.

To overcome these challenges, we develop a marginalized two-part (MTP) model for longitudinal semicontinuous data that yields more interpretable effect estimates when the primary focus is to estimate covariate effects on the average expenditures among the entire population of users and non-users. This model maintains many of the important features of conventional

longitudinal two-part models, such as capturing zero-inflation and skewness and accounting for clustering among observations through random effects, but allows investigators to examine covariate effects on the overall mean, a target of primary interest in many applications. We fit this model using a Bayesian approach, which can accommodate complex random effect structures in a computationally tractable manner using standard software such as SAS (SAS Institute, Cary, NC). We illustrate this approach by evaluating the effect of a copay increase on health care expenditures in the Veterans Affairs (VA) health care system over a four-year period.

The rest of the paper is organized as follows: Section 3.2 briefly reviews the conventional two-part model used with longitudinal semicontinuous data. Section 3.3 introduces the marginalized two-part model for longitudinal data, while Section 3.4 describes its computation in the Bayesian framework. Section 3.5 applies the approach to the VA study, and Section 3.6 provides a discussion and points to areas for future work.

3.2 Conventional Two-Part Model for Longitudinal Data

We begin with a review of the conventional two-part model. This model is commonly expressed as a two-part mixture of a point mass at zero and a distribution with positive support. The contribution of a given observation to this model is given by:

$$f(y_{ij}) = (1-\pi_{ij})^{1_{(y_{ij}=0)}} \times [\pi_{ij} g(y_{ij}|y_{ij}>0;\mu_{ij},\omega,\kappa)]^{1_{(y_{ij}>0)}} \qquad (3.1)$$
$$y_{ij} \geq 0, \quad j = 1,\ldots,n_i, \quad i = 1,\ldots,n,$$

where y_{ij} is the semicontinuous non-negative response for individual i at time j, $\pi_{ij} = \Pr(Y_{ij} > 0)$, μ_{ij} is the observation-level location parameter, ω is a scale parameter, κ is a skewness parameter, $1_{(\cdot)}$ is the indicator function, n_i is the number of responses observed for individual i, and n is the number of individuals. Any continuous density with positive support can be used for $g(y_{ij}|y_{ij}>0)$. For example, the log-skew-normal (LSN) distribution offers a flexible choice for $g(\cdot)$ that accommodates skewness on the log scale and includes the log-normal

distribution as a special case when $\kappa = 0$.

In the longitudinal setting, the probability and the location parameter, π_{ij} and μ_{ij}, are typically modeled in terms of fixed covariates and random effects as follows:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \boldsymbol{x}'_{1ij}\boldsymbol{\alpha} + \boldsymbol{z}'_{1ij}\boldsymbol{a}_i \quad \text{and} \\ \mu_{ij} &= \boldsymbol{x}'_{2ij}\boldsymbol{\gamma} + \boldsymbol{z}'_{2ij}\boldsymbol{c}_i, \end{aligned} \tag{3.2}$$

where $\pi_{ij} = \Pr(Y_{ij} > 0|\boldsymbol{b}_i)$, μ_{ij} is the location parameter for subject i at time j, $\boldsymbol{a}_i$ and $\boldsymbol{c}_i$ are random effects for the binary and continuous parts of the model, respectively; $\boldsymbol{b}'_i = (\boldsymbol{a}'_i, \boldsymbol{c}'_i)$ is assumed to follow a multivariate normal distribution; $\boldsymbol{x}_{1ij}$ and $\boldsymbol{x}_{2ij}$ are the vectors of the fixed effect covariates for subject i at time j and $\boldsymbol{z}_{1ij}$ and $\boldsymbol{z}_{2ij}$ are the corresponding vectors of the random effect covariates in the binary and continuous parts of the model, respectively. The fixed effect covariates, $\boldsymbol{x}_{1ij}$ and $\boldsymbol{x}_{2ij}$, are often chosen to be identical so that $\boldsymbol{x}_{1ij} = \boldsymbol{x}_{2ij} = \boldsymbol{x}_{ij}$. Note that while the binary part is commonly modeled with a logit link, a probit model could also be used.

Under this parameterization, γ_k, the kth regression coefficient in the continuous part of the model, represents the location shift in $\ln(Y_{ij}|Y_{ij} > 0)$ corresponding to a one-unit increase in the kth element of $\boldsymbol{x}_{2ij}$. In the simplest case, when $g(\cdot)$ in (3.1) is assumed to be lognormal, γ_k represents the shift in the subject-specific mean of $\ln(Y_{ij}|Y_{ij} > 0)$. If the more flexible LSN distribution is chosen, interpretation is further complicated as μ_{ij} is the location parameter rather than the mean; with either density, transformation back to the original scale is not straightforward. As a result, γ_k lacks a pragmatic interpretation in many settings.

Often, greater interest lies in estimating the effect of covariates on the marginal mean of Y_{ij} on the original (unlogged) scale for the combined population of users and non-users. For example, investigators may want to examine the association between a treatment and average cost among all patients, including health care users and non-users. As discussed in Smith et al. (2014), estimation of this effect is challenging under the conventional two-part model. Under the conventional model, the estimate of the treatment effect is conditional on fixed

values of the remaining covariates and will thus vary depending on values of these covariates. Moreoever, when random effects are included, point estimates of covariate effects on the overall mean require averaging over the random effects via numerical integration techniques (Liu et al. 2010; Tom et al. 2013). Further, to obtain confidence intervals on these estimates, resampling techniques such as bootstrapping must be employed (Liu et al. 2010). Because of the computational burden required to obtain inference on the covariate effects on the overall mean, some have considered population-average models that omit the random effects (Lu et al. 2004), inducing the assumption of independence between the binary and continuous parts, while adjusting the standard errors for correlation among repeated measures within each of the two parts separately. Ignoring the cross-part correlation, however, has been shown to induce bias in parameter estimates (Albert 2005; Su et al. 2009; Liu et al. 2010), leaving this as an unappealing solution. Thus, there remains the need for a flexible two-part model that accommodates dependence between components while providing an interpretable parameterization of the marginal mean in longitudinal studies.

3.3 Marginalized Two-Part Longitudinal Model

3.3.1 Model Specification

To overcome the limitations posed by the conventional two-part model, we propose a new marginalized two-part (MTP) longitudinal model that directly parameterizes the effect of covariates on the marginal mean of Y_{ij}, extending the MTP model developed in Smith et al. (2014) for cross-sectional semicontinuous data. The MTP model has the same two-part structure as the conventional model given in (3.1), but rather than parameterizing the model in terms of μ_{ij}, the log-scale location parameter of the conditionally positive values, we parameterize the model in terms of $\nu_{ij} = \mathrm{E}(Y_{ij}|\boldsymbol{b}_i)$, the overall mean among users and non-users combined. Thus, we have

$$\mathrm{logit}(\pi_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\alpha} + \boldsymbol{z}'_{1ij}\boldsymbol{a}_i \quad \text{and}$$

$$\ln(\nu_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta} + \boldsymbol{z}'_{2ij}\boldsymbol{d}_i \quad \text{or equivalently,} \tag{3.3}$$

$$\nu_{ij} = \mathrm{E}(Y_{ij}|\boldsymbol{b}_i) = \exp\left(\boldsymbol{x}'_{2ij}\boldsymbol{\beta} + \boldsymbol{z}'_{2ij}\boldsymbol{d}_i\right).$$

In contrast to $\boldsymbol{\gamma}$ and $\boldsymbol{c}_i$ in the conventional two-part model of (3.2), which represent effects on the subpopulation of users, $\boldsymbol{\beta}$ and $\boldsymbol{d}_i$ represent the fixed and random effects on the overall mean of combined users and non-users. Allowing $\boldsymbol{b}_i$ to now take $\boldsymbol{b}'_i = (\boldsymbol{a}'_i,\ \boldsymbol{d}'_i)$, the random effects are assumed to jointly follow a multivariate normal distribution, inducing cross-part correlation as in the conventional two-part model:

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{a}_i \\ \boldsymbol{d}_i \end{pmatrix} \sim N\left(\boldsymbol{0}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ad} \\ \boldsymbol{\Sigma}'_{ad} & \boldsymbol{\Sigma}_{dd} \end{bmatrix}\right). \tag{3.4}$$

Under the MTP model parameterization, β_k is the increment in the log of the overall mean, $\mathrm{E}(Y_{ij}|\boldsymbol{b}_i)$, corresponding to a unit increase in the kth covariate, x_{2kij}. This is in contrast to γ_k in equation (3.2) which represents a location shift in $\ln(Y_{ij}|Y_{ij} > 0)$. Thus, $\exp(\beta_k)$ represents the multiplicative effect of a unit increase in the kth covariate on the mean $\mathrm{E}(Y_{ij}|\boldsymbol{b}_i)$ for the entire population, including the users and non-users. Because this relationship is dependent on the parameterization and not specific to the distribution chosen for $g(\cdot)$ in (3.1), this interpretation remains the same regardless of the density chosen for the positive values.

3.3.2 Subject-Specific and Population Average Interpretations

Because random effects are included in the model, parameters naturally take a subject-specific interpretation. However, investigators are often interested in the effect of a covariate on population average outcomes as opposed to subject-specific ones. While a population average interpretation is often also referred to as a "marginal" interpretation, for clarity we reserve "marginal" in this chapter to refer to marginalizing over the populations of users and non-users. As detailed in Appendix B, under the MTP, the population average marginal mean

$\mathrm{E}(Y_{ij})$ is expressed as

$$\mathrm{E}(Y_{ij}) = \exp\left(\boldsymbol{x}'_{2ij}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}'_{2ij}\boldsymbol{\Sigma}_{dd}\boldsymbol{z}_{2ij}\right). \tag{3.5}$$

It is therefore a simple computation to estimate both subject-specific and population average marginal means, and functions thereof, under the MTP model parameterization. Specifically, differences in expected means between groups or ratios of such means are easily estimated from standard model output. Additionally, for covariates not included in $\boldsymbol{z}_{2ij}$, the regression effects take both subject-specific and population average interpretations. Specifically, for x_{2kij} not included in $\boldsymbol{z}_{2ij}$, the ratio of the population average means is given by:

$$\begin{aligned}\frac{\mathrm{E}(Y_{ij}|x_{2kij} = l+1, \boldsymbol{x}_{2(-k)ij})}{\mathrm{E}(Y_{ij}|x_{2kij} = l, \boldsymbol{x}_{2(-k)ij})} &= \frac{\exp\left(\boldsymbol{x}_{2(-k)ij}\boldsymbol{\beta}_{(-k)} + \beta_k \cdot (l+1) + \frac{1}{2}\left(\boldsymbol{z}'_{2ij}\boldsymbol{\Sigma}_{dd}\boldsymbol{z}_{2ij}\right)\right]}{\exp\left[\boldsymbol{x}_{2(-k)ij}\boldsymbol{\beta}_{(-k)} + \beta_k \cdot l + \frac{1}{2}\left(\boldsymbol{z}'_{2ij}\boldsymbol{\Sigma}_{dd}\boldsymbol{z}_{2ij}\right)\right]} \\ &= \exp(\beta_k),\end{aligned}$$

where $\boldsymbol{x}_{2(-k)ij}$ is $\boldsymbol{x}_{2ij}$ with the k covariate removed and $\boldsymbol{\beta}_{(-k)}$ is $\boldsymbol{\beta}$ with β_k removed. Details are provided in Appendix B. Under the MTP model, then, $\exp(\beta_k)$ has a dual interpretation as both a subject-specific and population average multiplicative effect on the marginal mean per unit increase in the kth covariate so long as that covariate is not included as a random effect. In many cases, investigators are interested in the effect of a binary covariate indicating receipt of some form of treatment or intervention, and this covariate is most often not included as a random effect, creating a dual interpretation for treatment effects commonly estimated by two-part models. In the special case when a random intercept alone is included in the overall mean specification, all covariate effects can be interpreted as population average effects.

3.4 Parameter Estimation, Computation, and Model Evaluation

Maximum likelihood estimation for the MTP model can present significant computational challenges, particularly when higher dimensional random effects are included in the model.

Our experience suggests that maximum likelihood estimation using an adaptive Gaussian quadrature often fails to converge for models that include random slopes. Even when models converge, run times can exceed several hours or even days for moderately large datasets (e.g., 24,000 individuals). We therefore adopt a fully Bayesian approach, which we have found to improve computational tractability relative to maximum likelihood and to accommodate models with correlated multidimensional random effects for both components of the MTP model.

To facilitate posterior inference, we make use of the likelihood in equation (3.1) by re-expressing the location parameter μ_{ij} in terms of ν_{ij} given in equation (3.3). For example, suppose we consider the LSN distribution for $g(\cdot)$ in equation (3.1), which takes the log-normal density as a special case while allowing greater flexibility (Azzalini 1985; Chai and Bailey 2008). The LSN density is given by

$$g(y_{ij}|y_{ij} > 0; \mu_{ij}, \omega, \kappa, \boldsymbol{b}_i) = \frac{2}{\omega y_{ij}} \phi\left(\frac{\ln y_{ij} - \mu_{ij}}{\omega}\right) \Phi\left(\frac{\kappa}{\omega}(\ln y_{ij} - \mu_{ij})\right),$$

where μ_{ij} is a subject-specific location parameter, $\omega > 0$ represents the scale parameter, and κ represents the shape parameter, all on the log scale; $\phi(\cdot)$ is the probability density function and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal density. In previous work (Smith et al. 2014), we found that the LSN density displayed better properties and more appropriately accounted for skewness commonly observed in semicontinuous data than the log-normal distribution. Using this density, the overall marginal subject-specific mean is given by

$$\mathrm{E}(Y_{ij}|\boldsymbol{b}_i) = \nu_{ij} = 2\pi_{ij} \exp\left(\mu_{ij} + \frac{\omega^2}{2}\right)\Phi(\omega\delta), \tag{3.6}$$

where $\delta = \frac{\kappa}{\sqrt{1+\kappa^2}}$, providing the link between μ_{ij} and ν_{ij}. The likelihood for the MTP model using the LSN distribution is in turn expressed as

$f(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \omega^2, \kappa, \boldsymbol{\Sigma}) =$

$$\prod_{i=1}^{n}\int_{-\infty}^{\infty}\left[\prod_{j=1}^{n_i}(1-\pi_{ij})^{1_{(y_{ij}=0)}}\left\{\frac{2\pi_{ij}}{\omega y_{ij}}\phi\left(\frac{\ln y_{ij}-\mu_{ij}}{\omega}\right)\Phi\left(\frac{\kappa}{\omega}(\ln y_{ij}-\mu_{ij})\right)\right\}^{1_{(y_{ij}>0)}}\right]N(\boldsymbol{b}_i;\mathbf{0},\boldsymbol{\Sigma})\mathrm{d}\boldsymbol{b}_i. \tag{3.7}$$

In order to utilize this likelihood for estimation of the MTP model, we solve equation (3.6) for μ_{ij} as a function of ν_{ij}, obtaining

$$\begin{aligned}\mu_{ij} &= \ln\nu_{ij} - \ln 2 - \ln\pi_{ij} - \ln\left[\Phi(\omega\delta)\right] - \frac{\omega^2}{2} \\ &= \boldsymbol{x}'_{2ij}\boldsymbol{\beta} + \boldsymbol{z}'_{2ij}\boldsymbol{d}_i - \ln 2 - \ln\pi_{ij} - \ln\left[\Phi(\omega\delta)\right] - \frac{\omega^2}{2}.\end{aligned} \tag{3.8}$$

We then plug this expression for μ_{ij} into (3.7) to proceed with model estimation.

Assuming prior independence, the joint posterior distribution of the parameters is given by

$$p(\boldsymbol{\alpha},\boldsymbol{\beta},\omega^2,\kappa,\boldsymbol{\Sigma}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\alpha},\boldsymbol{\beta},\omega^2,\kappa,\boldsymbol{\Sigma}) \times p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\omega^2)p(\kappa)p(\boldsymbol{\Sigma}), \tag{3.9}$$

where $f(\boldsymbol{y}|\boldsymbol{\alpha},\boldsymbol{\beta},\omega^2,\kappa,\boldsymbol{\Sigma})$ is given in equation (3.7) with μ_{ij} parameterized as in (3.8), and the $p(\cdot)$'s denote prior distributions for the respective parameters. To complete the Bayesian specification, we assign proper but diffuse priors to model parameters. For the fixed effects, we assume $\boldsymbol{\alpha},\boldsymbol{\beta} \sim N(\mathbf{0},\boldsymbol{\Sigma}_0)$ where $\boldsymbol{\Sigma}_0$ is a diagonal matrix with diagonal elements of 1000. We assume $\boldsymbol{\Sigma}$, the covariance of the random effects, follows an inverse-Wishart IW$(q,\boldsymbol{I})$ prior distribution with q degrees of freedom, where $\boldsymbol{I}$ is the $q \times q$ identity matrix. Scale and shape parameters, ω^2 and κ are assumed to follow an inverse-Gamma IG$(0.001, 0.001)$ and a Uniform(-10, 10) prior distribution, respectively. In our experience, the range from -10 to 10 for κ is sufficient to capture any degree of skewness that is likely to be observed in practice.

For posterior computation we use Markov chain Monte Carlo (MCMC), which iteratively draws samples of the model parameters. At convergence, the chain achieves a stationary distribution that is the joint posterior distribution of the model parameters. The MCMC computation can be implemented conveniently in standard software such as SAS `PROC MCMC`,

which employs a hybrid of random walk Metropolis-Hastings steps and conjugate Gibbs updates. Appendix C provides `PROC MCMC` code for fitting the LSN MTP model.

We assess model fit, and in particular, the complexity of the random effects specification needed, via model selection using the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002). Letting $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \omega^2, \kappa, \boldsymbol{\Sigma})'$ denote the collection of all model parameters, the DIC is defined as $\bar{D}(\boldsymbol{\theta}) + p_D$, where $\bar{D}(\boldsymbol{\theta})$ is considered a measure of the goodness of fit and p_D is a penalty for model complexity. As with other penalized selection criteria, lower DIC values indicate better model fit (Spiegelhalter et al. 2002).

3.5 Analysis of Change in VA Specialty Care Copayment

The Veterans Health Administration (VA) significantly increased its outpatient visit copayments in December 2001, with primary care copayments increasing from \$0 to \$15 per visit and specialty care copayments increasing from \$15 per visit to \$50. Reflecting changes in the health care market, this change created a natural experiment to examine the impact of copayments on health expenditures for outpatient specialty care. Veterans with sufficiently low income or with sufficient disability from military service are exempt from outpatient visit copayments, creating a control group that did not experience this increase in price. The data used to assess the impact of this copay change have been described previously (Maciejewski et al. 2010a; 2012a). Briefly, 51,503 veterans with hypertension who were diagnosed and prescribed an antihypertensive medication in 2000 at four VA Medical Centers (VAMCs) were identified. Veterans were then excluded if they: (i) were not alive during the entire study period (2000-2003) (n=7007); (ii) had a majority of their primary care visits outside of these four VAMCs (n=10,317); (iii) had an unknown military-service-connected disability needed to determine copayment exemption (n=47); or (iv) were hospitalized when the copayment increase went into effect or for more than one year during the study period (n=29).

Because the data were originally drawn for a study designed to assess the impact of a medication copayment increase in 2002, veterans were also excluded if they did not have a required history of medication fills (n=13,095) or their medication copayment exemption

status was unknown (n=13,068). These exclusions resulted in an sample of 7940 veterans with hypertension, including 4395 copay-exempt veterans and 3545 veterans required to pay copayments.

To reduce non-equivalence between the copay exempt and required cohorts, veterans in the two groups were matched via one-to-one propensity score matching without replacement using a modified version of the Parsons' nearest neighbor greedy matching algorithm (Parsons 2001). The propensity score model included age, sex, race, marital status, comorbidity burden in 2000 via the Diagnostic Cost Group (DCG) score, depression diagnosis in 2000, number of antihypertensive medications in 2000, VAMC, and ZIP code-level variables based on year 2000 census data of proportion of the population in each category of highest education level attained (less than high school, high school, college or higher) and mean per capita income. Matching resulted in an analytic sample of 1693 veterans exempt from copayments and 1693 veterans required to pay copayments who were well matched. Descriptive statistics for these two cohorts are presented in Table 3.1.

The expenditure outcome of interest included all outpatient expenditures for specialty care, identified using clinic identifiers in the VA administrative data. The annual VA expenditures for outpatient specialty visits were constructed for each patient in each observation year (2000-2003), two years prior to and two years after the copayment change. All expenditures were inflation adjusted to 2003 dollars using the medical component of the consumer price index.

The percentage of patients with zero expenditures ranged from 36% to 16% over the study period, yielding the need for a two-part model. To assess the effect of the copayment increase on specialty care expenditures, we fit the MTP model with identical explanatory variables in the binary and overall mean components: (i) an indicator of whether or not a veteran was required to pay copayments, (ii) fixed effects for each year with year 2000 as reference, and (iii) interactions between the year fixed effects and copay exemption status. Random intercepts and slopes were included in both components with a 4×4 covariance matrix. For comparison, we fit a reduced model that excluded the random slope in the binary part, resulting in a 3×3

covariance matrix for the random effects.

The DIC values for the full and reduced models were 167,798 and 168,235, respectively, indicating the better fit for the full model with random intercepts and slopes in both components. Posterior means and standard deviations of the parameters from both components of this model are presented in Table 3.2, and Figure 3.1 displays the model estimated means at each year with their 95% credible intervals (CIs). Model specification details and convergence diagnostics are presented in Appendix D.

Specialty care expenditures remained lower for those required to pay copays throughout the study period than for those exempt. After the copayment increased in December 2001, expenditures among the copay-required cohort decreased very slightly while the expenditures among the copay-exempt cohort continued to increase. Specifically, we can estimate the multiplicative difference in expenditures among those having to pay the copayment compared to those exempt in each year by exponentiating β_4 through β_7 in Table 3.2 for years 2000 through 2003, respectively. Note that because the copayment indicator and interactions were not included as random effects, these differences can be interpreted as both population average and subject-specific effects. Additionally, by computing the difference in means using equation (3.5) at each iteration of the MCMC chain, we can estimate the additive difference in population average means with 95% CIs. These results are shown in Table 3.3.

Those required to pay copayments had 0.71 times the expenditures (95% CI: [0.61, 0.81]) of those exempt in 2000, prior to the copayment increase. Two years after the increase, those required to pay copayments had 0.51 times the expenditures (95% CI: [0.45, 0.58]), suggesting a notable impact of the copayment increase on mean outpatient specialty care expenditures. Similarly, those required to pay copayments had on average $363 lower expenditures (95% CI: [-$519, -$219]) in 2000 than those exempt, but had on average $803 lower expenditures in 2003 (95% CI: [-$988, -$639]).

3.6 Conclusion

We proposed a marginalized two-part model for longitudinal semicontinuous data that allows investigators to obtain the effect of covariates on the overall population mean. Our model directly parameterizes the covariates in terms of the population mean while still appropriately accounting for the excess number of zeros and correlation between the two model components. It allows for estimation of the overall population average mean, in addition to subject-specific means, and many covariates take a dual population average and subject-specific interpretation. The proposed Bayesian inferential approach can easily accommodate complex random effect structures and estimate credible intervals for quantities of practical interest, such as differences in mean expenditures.

Using the MTP model, we estimated that the requirement to pay a VA copayment for specialty care outpatient visits was associated with lower specialty care outpatient expenditures compared to a control group that was exempt from copayments, and that an increase in this copayment was associated with a larger difference in expenditures. Specifically, the mean difference in outpatient specialty care expenditures among those required to pay compared to those exempt increased from $363 in 2000, two years prior to the copayment increase, to $803 in 2003, two years after the copayment increase. In contrast to the conventional two-part model, which provides estimates of the effect of a policy change separately on the probability of incurring expenditures and on the log of expenditures given that are incurred, these MTP results are directly interpretable and useful for informing policy decisions.

These results suggest that copayment increases can significantly reduce demand for outpatient specialty care, which contributes to the continued escalation in health expenditures in the United States. As health expenditures continue to increase, the MTP model will be useful for examining a variety of interventions that may be utilized to restrain the continued growth. Further, the MTP model could also be more generally useful in estimating aggregate health expenditures in a population, which private insurers must do annually for premium rate setting and government agencies must do when preparing budget requests for Congress.

Future directions for the MTP model could include extensions to spatially correlated data or multi-level hierarchical models and applications to other fields, such as substance abuse or psychometric research. The MTP model could also be extended to incorporate latent classes for examining heterogeneous treatment effects across patient subpopulations.

In short, the proposed longitudinal MTP model provides a straightforward method for estimating covariate effects on the marginal mean of the population as a whole. Such effects are challenging to estimate in conventional two-part models. Further, many other quantities, such as the overall mean and differences in means, can be easily estimated on the original scale of the data. As such, the MTP model simplifies economic evaluations that are increasingly critical to understanding the return on investment of new interventions, policies and programs.

Figure 3.1: Model estimated mean expenditures and 95% credible intervals (shaded regions) for the outpatient specialty care analysis

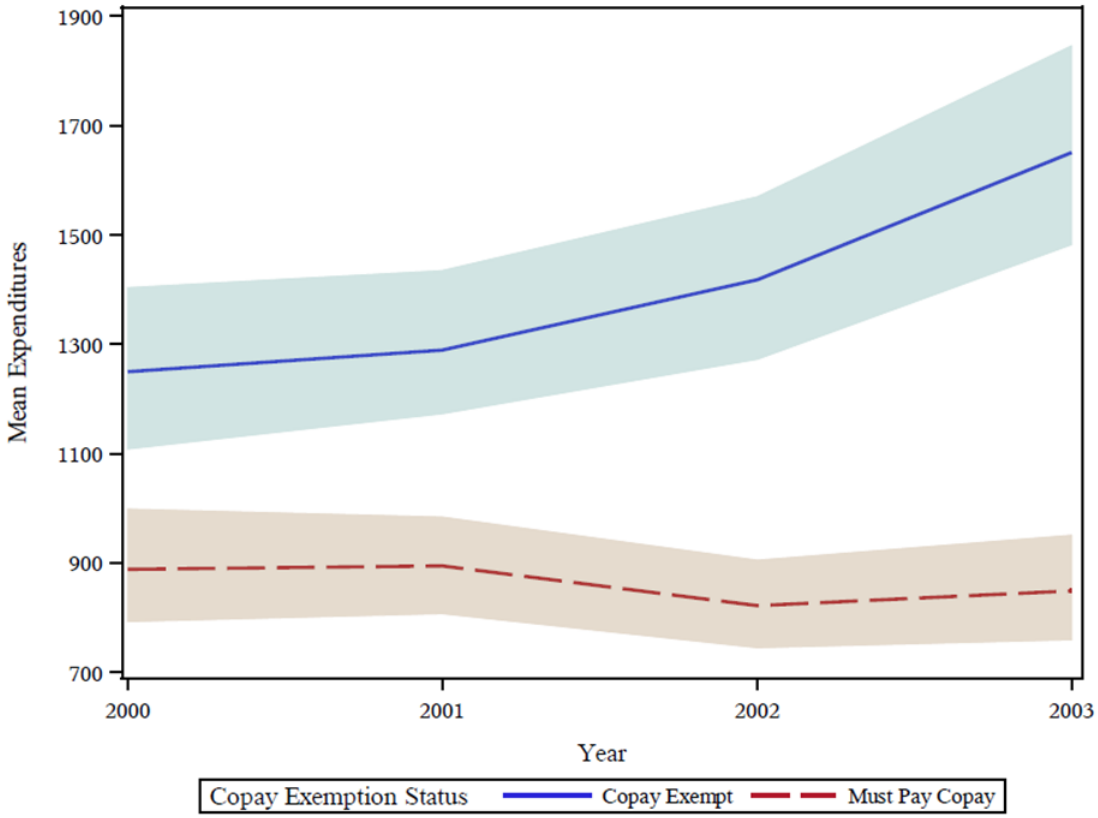


Table 3.1: Descriptive statistics of the matched cohorts in the outpatient specialty care copay study

	Copayment Status	
	Exempt $n = 1693$	Nonexempt $n = 1693$
	Mean (SD)	
Age	65.8 (10.9)	66.1 (10.8)
DCG score in 2000	0.76 (1.16)	0.75 (1.13)
Baseline number of antihypertensive medications	7.3 (4.3)	7.1 (4.0)
Proportion with < high school education in ZIP code	18 (11)	18 (10)
Proportion with high school education in ZIP code	53 (10)	53 (10)
Proportion with college education in ZIP code	28 (16)	29 (16)
Mean per capita income in ZIP code in $10,000	5.19 (1.70)	5.21 (1.70)
	Percent	
Male	97.3	97.3
White	65.3	65.8
Nonwhite	13.6	13.6
Unknown race	21.1	20.6
Married	69.9	69.1
Depression diagnosis at baseline (%)	3.8	3.2
VAMC A	15.6	16.0
VAMC B	25.1	23.6
VAMC C	36.0	36.7
VAMC D	23.3	23.7

Table 3.2: Posterior means and 95% credible intervals of MTP model parameters

	Parameter	Posterior Mean	95% Credible Interval
Binary Component			
Intercept	α_0	2.13	(1.89, 2.35)
Year 2001	α_1	0.31	(0.09, 0.53)
Year 2002	α_2	0.83	(0.56, 1.12)
Year 2003	α_3	0.65	(0.30, 0.99)
Must Pay Copay	α_4	-1.06	(-1.33, -0.78)
Year 2001 × Must Pay	α_5	0.005	(-0.28, 0.30)
Year 2002 × Must Pay	α_6	-0.15	(-0.49, 0.15)
Year 2003 × Must Pay	α_7	-0.34	(-0.68, -0.0004)
Overall Mean Component			
Intercept	β_0	6.24	(6.15, 6.33)
Year 2001	β_1	0.13	(0.05, 0.22)
Year 2002	β_2	0.23	(0.13, 0.32)
Year 2003	β_3	0.28	(0.17, 0.37)
Copay Required	β_4	-0.34	(-0.49, -0.21)
Year 2001 × Must Pay	β_5	-0.02	(-0.15, 0.10)
Year 2002 × Must Pay	β_6	-0.20	(-0.33, -0.08)
Year 2003 × Must Pay	β_7	-0.32	(-0.46, -0.18)
Shape and Scale Parameters			
Shape	κ	-0.95	(-1.22, -0.61)
Scale	ω^2	1.54	(1.30, 1.75)
Covariance of Random Effects			
Variance of Binary Random Intercept	σ^2_{11}	5.18	(4.19, 6.21)
Variance of Binary Random Slope	σ^2_{22}	0.46	(0.31, 0.62)
Variance of Overall Random Intercept	σ^2_{33}	1.77	(1.59, 1.98)
Variance of Overall Random Slope	σ^2_{44}	0.10	(0.08, 0.13)
Covariance Parameters:	σ_{12}	-0.30	(-0.59, -0.01)
	σ_{13}	2.51	(2.18, 2.84)
	σ_{14}	-0.18	(-0.28, -0.07)
	σ_{23}	-0.0004	(-0.12, 0.14)
	σ_{24}	0.11	(0.06, 0.16)
	σ_{34}	-0.15	(-0.21, -0.10)

Table 3.3: Model estimated effects of copayment requirement

	Multiplicative Effect		Additive Effect	
Year	Mathematical Expression	Estimated Effect (95% CI)	Mathematical Expression	Estimated Effect (95% CI)
2000	e^{β_4}	0.71 (0.61, 0.81)	$E(Y_{i,2000}\|\text{Pay}=1) - E(Y_{i,2000}\|\text{Pay}=0)$	-$363 (-$519, -$219)
2001	e^{β_5}	0.69 (0.61, 0.78)	$E(Y_{i,2001}\|\text{Pay}=1) - E(Y_{i,2001}\|\text{Pay}=0)$	-$396 (-$549, -$265)
2002	e^{β_6}	0.58 (0.51, 0.65)	$E(Y_{i,2002}\|\text{Pay}=1) - E(Y_{i,2002}\|\text{Pay}=0)$	-$597 (-$746, -$455)
2003	e^{β_7}	0.51 (0.45, 0.58)	$E(Y_{i,2003}\|\text{Pay}=1) - E(Y_{i,2003}\|\text{Pay}=0)$	-$803 (-$988, -$639)

CHAPTER 4: COMPARISON OF ONE-PART MODELS AND A TWO-PART MARGINALIZED MODEL FOR THE ANALYSIS OF HEALTH CARE EXPENDITURES

4.1 Introduction

There are a number of analytic challenges associated with the analysis of health care expenditures. They are often characterized by two defining features: a portion of the sample who are non-users with zero expenditures, and a highly skewed distribution of expenditures among those who are users. Data with such features are often deemed "semicontinuous" to describe the mixture of the discrete point mass at zero with the skewed distribution of positive values. They are often thought of as arising from two distinct stochastic processes: one governing the occurrence of zeros and the second determining the observed value conditional on it being a nonzero response. The first process is commonly referred to as the "binary" part of the data, while the second is often termed the "continuous" part.

To accommodate these two processes, analysts often consider two-part models. Because they explicitly accommodate both data generating processes, two-part models can be an ideal choice for modeling semicontinuous data. Most commonly, the binary part is modeled via logistic regression and the continuous component via a log-normal model. However, because the log-normal distribution imposes a sometimes unrealistic condition of symmetry on the log-scale, alternative distributions such as the log-skew-normal or generalized gamma have recently been proposed for the continuous part in an effort to relax these assumptions (Azzalini 1985; Chai and Bailey 2008; Manning et al. 2005; Liu et al. 2010). When adjusting for covariates, these models typically include one set of parameters for the binary response and a second set for the continuous component conditional on a positive response. In particular, covariates in the second, or continuous, part are interpreted conditionally upon having ob-

served a positive outcome. Attempts to combine these two parts to form the overall marginal mean effect of any covariate relies on specifying values for each of the other covariates in the model, and therefore varies depending on those values. As such, it is generally challenging to obtain a straightforward interpretation of covariate effects on the marginal mean in two-part models.

In many cases, however, investigators' main interest lies in examining effects on the marginal mean in order to draw conclusions about the impact of predictors on the population as a whole. To accomplish this, "one-part" models provide an attractive alternative. One-part models incorporate both the zero and positively continuous values as arising from the same stochastic process rather than explicitly accounting for the point mass at zero, and in doing so, they permit interpretation of covariate effects on the overall mean. These models typically take one of two general forms. In one form, a small constant is added to the outcome to ensure all values are positive and the outcome is then transformed to minimize skewness. Most commonly, a linear model for the log transformed outcome is used. Alternatively, a generalized linear model (GLM) can be utilized, often with a log link, to avoid transformation and the need to add a constant to all values. Further, these GLMs can be fit using quasilikelihood with empirical sandwich standard errors (Royall 1986; Kauermann and Carroll 2001), avoiding parametric assumptions. While these standard errors provide asymptotically valid inference even if the variance model is misspecified, their finite sample performance in the presence of many zero values has not been fully evaluated.

Recently, Smith et al. (2014) proposed a fully parametric marginalized two-part (MTP) modeling approach that specifies the same marginal mean model as a typical one-part GLM with log link while simultaneously accounting for the point mass at zero. Rather than parameterizing the model in terms of the mean of the transformed, conditionally positive outcomes in the second part, as in other two-part models, the MTP model parameterizes covariate effects directly on the overall mean, $\mathrm{E}(Y)$, on the untransformed scale. This allows parameter estimates to be interpreted as the multiplicative effect on the overall mean rather than on the conditional mean of only the positive outcomes. This approach also has the advan-

tage of separately providing estimates of covariate effects on the probability of incurring a positive-valued outcome, as in the first part of two-part models, as well as accounting for the zero-inflated and skewed nature of many semicontinuous outcomes. On the other hand, however, it relies on fully parametric assumptions, unlike GLMs fit with quasilikelihood.

Because these GLMs rely on fewer assumptions, it may be natural to question whether one-part models fit to semicontinuous data perform better than MTP models in terms of bias and precision when interest lies in marginal inferences on the overall mean. Duan et al. (1983), Diehr et al. (1999), Madden et al. (2000), and Buntin and Zaslavsky (2004) have each compared the performance of "conventional" two-part models with various one-part models. In each case, the models were assessed using real datasets and performance was determined using a combination of goodness of fit criteria and predictive accuracy. Conclusions were mixed, with one-part models performing equally well or better on some datasets and two-part models exhibiting better performance on others.

In particular, Buntin and Zaslavsky fit one-part GLMs with quasi-likelihood, suggesting that excess zeros in the data pose no problem when fitting a GLM due to the use of a link function rather than using a log-transformed outcome. They tested several one- and two-part models with the goal of predicting Medicare expenditures using a sample with 8.6% of individuals having zero expenditures and assessed each model's predictive ability via split-sample cross-validation. They concluded that, unless there is specific interest in separately modeling the probability of positive expenditures, researchers begin by fitting one-part models, with the caveat that this may lead to reduced efficiency in standard errors relative to correctly-specified parametric models. If the probability of positive expenditures were of specific interest, or if the researchers were unable to find a suitably fitting one-part model, they then suggested proceeding to examine two-part models.

While Buntin and Zaslavsky have provided the main comparison to date between the predictive abilities of one-part GLMs and conventional two-part models, more work is needed to assess model performance under the presence of a greater proportion of zeros as well as the ability of one- and two-part models to accurately estimate the effects of covariates. Previously,

comparing model-estimated covariate effects across one- and two-part models was not generally possible because the conventional two-part model separately specified the probability of a positive expenditure and the level of expenditure conditional on it being positive. The recent introduction of the MTP model, however, provides an analysis approach that that explicitly accounts for excess zeros without sacrificing the interpretability of covariate effects on the overall mean. The performance of the MTP model has not been examined in comparison to one-part models, nor has a formal simulation study comparing such one- and two-part models been conducted.

To further evaluate differences in these modeling approaches, we report simulation results on the performance of the MTP model and three different GLMs fit with quasilikelihood using empirical standard errors. This simulation design was motivated in part by an analysis to assess the impact of a behavioral weight loss program on health care expenditures in the year following enrollment, presented in Smith et al. (2014). We evaluate bias, test size, and coverage of nominal 95% confidence intervals under varying data generating mechanisms.

The remainder of this paper is laid out as follows. Section 4.2 briefly reviews the MTP model and GLMs fit with quasilikelihood, while Section 4.3 discusses the details of the simulations conducted. Section 4.4 shows the results of the simulations, and Section 4.5 provides a discussion of the implications of the results and points to areas for future research and investigation.

4.2 Models Compared

In this simulation study, we compare the MTP model developed in Smith et al. (2014) and GLMs fit with quasilikelihood. These models take the same mean structure, providing easily comparable quantities, and both fit the data on the original untransformed scale, so retransformation methods are not required. We begin with a brief review of the models considered.

4.2.1 MTP Model

For data consisting of independent observations, the generic form of a two-part model can be written as

$$f(y_i) = (1-\pi_i)^{1_{(y_i=0)}} \times [\pi_i g(y_i|y_i>0)]^{1_{(y_i>0)}}, \quad y_i \geq 0,\ i=1,\ldots,n, \tag{4.1}$$

where $\pi_i = \Pr(Y_i > 0)$, $1_{(\cdot)}$ is the indicator function, and $g(y_i|y_i>0)$ is any density function applicable to the positive values of Y_i. To obtain interpretable covariate effects on the marginal mean, Smith et al. (2014) proposed the MTP model that parameterizes the covariate effects directly in terms of the marginal mean, $\nu_i = \mathrm{E}(Y_i)$, on the original (i.e., untransformed) data scale. The MTP model specifies the linear predictors

$$\begin{aligned} \mathrm{logit}(\pi_i) &= \boldsymbol{z}_i'\boldsymbol{\alpha} \quad \text{and} \\ \mathrm{E}(Y_i) = \nu_i &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta}). \end{aligned} \tag{4.2}$$

Smith et al. developed this model with $g(y_i|y_i>0)$ taking either the log-normal or log-skew-normal (LSN) density. The LSN density relaxes the the log-normal density's assumption of log-scale normality through inclusion of a shape parameter, κ, allowing skewness on the log-scale, with the log-normal density taking the special case of $\kappa = 0$. In previous work (Smith et al. 2014), we found that the LSN density displayed better properties and more appropriately accounted for skewness commonly observed in semicontinuous data than the log-normal distribution. For this reason, we focus on the LSN MTP model here.

Using the linear predictors as in equation (4.2), the generic form of the two-part LSN model for independent data is given by:

$$f(y_i) = (1-\pi_i)^{1_{(y_i=0)}} \times [\pi_i \mathrm{LSN}(y_i;\xi_i,\omega,\kappa)]^{1_{(y_i>0)}}, \quad y_i \geq 0,\ i=1,\ldots,n, \tag{4.3}$$

where $\mathrm{LSN}(\cdot;\xi_i,\omega,\kappa)$ denotes LSN distribution with location parameter ξ_i, scale parameter

$\omega > 0$, and shape parameter κ, all on the log scale, given by

$$g(y_i|y_i > 0) = \frac{2}{\omega y_i}\phi\left(\frac{\ln y_i - \xi_i}{\omega}\right)\Phi\left(\frac{\kappa}{\omega}(\ln y_i - \xi_i)\right), \tag{4.4}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function, respectively, of the standard normal density. The marginal mean of Y_i is then given by:

$$\mathrm{E}(Y_i) = \nu_i = 2\pi_i \exp\left(\xi_i + \frac{\omega^2}{2}\right)\Phi(\omega\delta), \tag{4.5}$$

where $\delta = \frac{\kappa}{\sqrt{1+\kappa^2}}$. In order to re-express the LSN likelihood as a function of $\boldsymbol{\beta}$, we solve equation (4.5) for ξ_i in terms of $\boldsymbol{\beta}$:

$$\begin{aligned}\xi_i =& \ln \nu_i - \ln 2 - \ln \pi_i - \ln\left[\Phi(\omega\delta)\right] - \frac{\omega^2}{2} \\ =& \boldsymbol{x}_i'\boldsymbol{\beta} - \ln 2 - \ln \pi_i - \ln\left[\Phi(\omega\delta)\right] - \frac{\omega^2}{2}.\end{aligned}$$

After plugging this expression into equation (4.3) above, parameter estimates can be obtained using standard optimization routines such as Newton-Raphson or Fisher scoring. Model-predicted means and standard errors can also be easily obtained under this parameterization in a single step by estimating $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ at the desired values of the covariates. SAS code (SAS Institute, Cary, NC) implementing the MTP model using `PROC NLMIXED` is provided in Smith et al. (2014).

4.2.2 GLMs Fit with Quasilikelihood

GLMs fit using quasilikelihood require only the specification of the mean and variance, as opposed to the full distribution, making them an attractive alternative when assumptions regarding the underlying parametric distribution are questionable. Specifically, when using a log link as is most commonly specified for health care expenditures, the overall mean model

is given by

$$\mathrm{E}(Y_i) = \nu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \tag{4.6}$$

the same as specified in the MTP model. A commonly used family of variance functions is the power family, taking the form

$$\mathrm{Var}(Y_i) = \rho\nu_i^{\lambda} = \rho\exp(\boldsymbol{x}_i'\boldsymbol{\beta})^{\lambda}, \tag{4.7}$$

and methods have been proposed to assist in determining the optimal value of λ (Manning and Mullahy 2001; Park 1966; Basu and Rathouz 2005). Specifically, commonly used values include $\lambda = 0$, constant variance, $\lambda = 1$, variance proportional to the mean, and $\lambda = 2$, variance proportional to the square of the mean, or equivalently, standard deviation proportional to the mean. Empirical "sandwich" variance estimators (Royall 1986; Kauermann and Carroll 2001) are commonly paired with such GLMs, such that if the variance is misspecified, they yield valid inference under many conditions where the marginal mean model is correctly specified (Fitzmaurice et al. 2012). For this comparison, we utilize the empirical standard errors and fit GLMs with $\lambda = 0$, 1, and 2, or with constant variance, variance proportional to the mean, and standard deviation proportional to the mean. Such models are implementable in most standard statistical software packages.

4.3 Simulation Details

4.3.1 Mean Structure and Properties Examined

To evaluate the performance of the LSN MTP and the GLMs, we conducted a series of simulation studies motivated in part by the analysis of a behavioral weight loss program presented in Smith et al. (2014). That study evaluated the effect of a system-wide weight loss intervention (MOVE!) implemented by the Veterans Affairs (VA) health care system beginning in 2006 to address the high prevalence of obesity among VA patients (Kahwati et al.

2011). As part of that study, the total expenditures in the year following enrollment of 18,214 MOVE! enrollees were compared to those of 18,214 non-enrollees who were matched to the enrollees on sex, race (white or non-white), marital status (married or non-married), copay status (exempt vs. non-exempt), veterans integrated service network (VISN) of residence, BMI, and comorbidity burden, assessed via the 2002 diagnostic cost group (DCG) score. The goal of the analysis was to assess whether MOVE! enrollment was associated with a difference in total health care costs in the following year. With 17% of the MOVE! enrollees having zero expenditures in the year, results from one-part GLMs may have been unreliable, and use of the MTP model was therefore motivated.

Basing covariate distributions and parameter values on those of the MOVE! study, all simulated data scenarios considered here were generated assuming the following marginal mean structure:

$$\mathrm{E}(Y_i) = \nu_i = \exp(6 + 0.2x_{1i} - 0.01x_{2i} + 0.05x_{3i}), \tag{4.8}$$

where $x_{1i} \sim$ Bernoulli(0.5), $x_{2i} \sim N(0,1)$, and $x_{3i} \sim$ Pois(1). We considered three different scenarios for the distribution of the positive values of Y_i: (1) distributed as LSN with low log-scale skewness, (2) distributed as LSN with higher log-scale skewness, and (3) distributed as generalized gamma (GG). For each of these three scenarios, we considered data with approximately 20% zeros and approximately 40% zeros to assess the influence of the size of the discrete point mass on the performance of each model. Specifically, zeros were introduced in the Y_i's with probability π_i, where π_i was given by $\mathrm{logit}(\pi_i) = 3 - 4x_{1i} + 3.5x_{2i} + 2.5x_{3i}$ and $\mathrm{logit}(\pi_i) = 3 - 7x_{1i} + 5x_{2i} + 2x_{3i}$ to achieve approximately 20% and 40% zeros, respectively. For each of these six combinations of distributions and percentages of zeros, we evaluated datasets of sample sizes 200, 1,000, and 10,000 to assess the impact of sample size on model performance, resulting in a total of 18 simulations with 1,000 datasets each. In each case, the mean model of the GLMs and MTP models fit to the data were correctly specified as $\mathrm{E}(Y_i) = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})$.

To assess the performance of each model, we examined the mean bias, median bias, and

percent relative median bias of parameter estimates, as well as mean and median bias of model-predicted "total cost", the simulated outcome. Total cost bias was calculated as the average difference in the prediction of individuals' observations and their true theoretical means, based on their respective combination of covariates. We also examined coverage probability of nominal 95% Wald-type confidence intervals for each parameter as well as total cost predictions. For each of the 18 scenarios, we then re-generated data with $\beta_1 = 0$ to mimic a null hypothesis of no treatment effect for the binary variable x_{1i} in order to evaluate Type I error rates for each model at a nominal 0.05 significance level.

4.3.2 Simulation 1: Log-Skew-Normal Data

In the first set of simulations, we assumed the positive values of Y_i followed the LSN density shown in equation (4.4). Thus, in this simulation, the parametric assumptions of the MTP model were met. We set the scale parameter, ω, at 1.2 and set the log-scale skewness parameter κ at 0.5 and 5.0 for the low log-scale skewness and high log-scale skewness simulations, respectively.

4.3.3 Simulation 2: Generalized Gamma Data

In the second set of simulations, we investigated the performance of the MTP relative to that of the one-part GLMs when the parametric distributional assumptions were not met. The generalized gamma is a flexible, three-parameter distribution that takes as special cases the standard gamma, inverse gamma, Weibull, and log-normal distributions (Manning et al. 2005; Liu et al. 2010). The density is given by

$$f(y_i; \kappa, \mu_i, \sigma) = \frac{\eta^\eta}{\sigma y_i \Gamma(\eta)\sqrt{\eta}} \exp\left[u\sqrt{\eta} - \eta \exp(|\kappa|u)\right],$$

where $\eta = |\kappa|^{-2}$, $u = \text{sign}(\kappa)\left(\log(y_i) - \mu_i\right)/\sigma$, μ_i is the location parameter, $\sigma > 0$ is the scale parameter, and κ is the shape parameter. As in Simulation 1, we set the scale parameter, σ, at 1.2, and we and set the shape parameter, κ, at 0.63 based on the analysis from Liu et al.

(2010). Computational details from both simulations are provided in Appendix E.

4.4 Simulation Results

4.4.1 Log-Skew-Normal Results

Descriptive statistics on the six datasets generated from the LSN distribution with lower log-scale skewness are shown in Table 4.1, and median bias from the models fit on each of these datasets is shown in Table 4.2. Results with higher log-scale skewness were similar and are shown in Appendix E, along with mean bias and percent relative median bias. The MTP model provided the least biased estimates under all scenarios, which is expected given the parametric assumptions of the model were met. Among all models, bias generally decreased with sample size, and among the GLMs, was noticeably larger when the data had 40% zeros as opposed to 20%. In particular, β_1, the treatment effect of main interest, was negatively biased under all of the GLMs. With 40% zeros, the negative bias increased such that, for sample sizes of 200 and 1,000, the GLMs were on average producing negative treatment effect estimates instead of positive ones.

Table 4.3 shows coverage probabilities of the 95% Wald-type confidence intervals from the models fit to each of the LSN generated datasets. The MTP model maintained approximately 0.95 coverage probability under all scenarios. Even with empirical standard errors, modest reductions in coverage probability were seen for the GLMs with 20% zeros, with coverage ranging from 0.74 to 0.93. With 40% zeros, coverage dropped significantly for the GLMs, particularly for the effects of covariates. In particular, coverage for β_1, the treatment effect, ranged from 0.48 to 0.75 for the GLMs with 40% zeros. No clear pattern was seen in the coverage probability with increasing sample size.

4.4.2 Generalized Gamma Results

Descriptive statistics on the six datasets generated from the GG distribution are shown in Table 4.4, and median bias from the models fit on each of these datasets is shown in Table

4.5. Under this scenario, when the parametric assumptions of the MTP model were no longer met, the MTP model appeared to incur slightly more bias in estimating the intercept, β_0, and subsequently, in total cost prediction. Notably, the bias in the intercept and total cost prediction did not improve with increased sample size. For the estimation of covariate effects, however, bias remained low for the MTP model regardless of sample size or percentage of zeros. The GLMs again performed much better with 20% zeros than with 40%, and the bias incurred appeared to decrease with sample size. Even with a sample of 10,000, however, the estimate of treatment effect under the GLMs with 40% zeros was strongly negatively biased, and with the smaller sample sizes, often resulted in estimates of treatment effect that were in the wrong direction.

Similar trends were seen in the coverage probabilities shown in Table 4.6. Coverage probabilities for the intercept and total cost prediction under the MTP model dropped to as low as 0.35 with a sample size of 10,000. Coverage for the covariate effect parameters, however, remained close to 0.95 under the MTP model regardless of sample size or percentage of zeros. Similar to the results using the LSN data, the GLMs showed a modest reduction in coverage with 20% zeros, with values ranging from 0.74 to 0.92. With 40% zeros, however, coverage for the GLMs dropped significantly for all parameters. In particular, coverage for β_1, the treatment effect, ranged from 0.48 to 0.74 in this scenario. Coverage for total cost prediction with 40% zeros was higher for the MTP under the smaller sample sizes of 200 and 1,000, but with 10,000 subjects, the MTP model coverage of total cost prediction dropped substantially and the GLMs provided higher coverage. None of the models provided particularly good coverage with 40% zeros and 10,000 subjects, with the highest coverage probability being 0.70.

4.4.3 Type I Error Rates

Type I error rates from each of the models re-run on data simulated with $\beta_1 = 0$ under the LSN distribution with low log-scale skewness and the GG distibution are shown in Table 4.7. Results were again similar for the LSN distribution with high log-scale skewness and

are included in Appendix E. Type I error rates remained close to 0.05 for the MTP model under all scenarios, while type I errors remained at least somewhat inflated under almost all scenarios for the GLMs. When the data contained 20% zeros, the GLM type I error rates ranged from 0.07 to 0.16. With 40% zeros, they ranged from 0.26 to 0.52. Type I errors seemed to generally decrease with increasing sample size for the GLMs, but particularly with 40% zeros, rates remained significantly higher than the nominal 0.05 significance level at all sample sizes examined.

4.5 Discussion

Results suggest one-part GLMs fit with quasilikelihood may not in general provide good alternatives for modeling datasets consisting of a significant proportion of zeros. Even with 20% zeros, some bias and decreased coverage was seen, although this improved with larger sample sizes. When the distributional assumptions of the MTP were not met, however, the GLMs provided better predictive accuracy for the total cost outcome, especially when the percentage of zeros was lower and the sample size was large.

As most health care expenditure datasets are quite complex, there does not appear to be a single easy solution to finding easily interpretable models that also provide good fit. Analysts must carefully consider both the properties of their datasets as well as the goals of their analysis. Our results suggest that if the primary goal is to estimate the effect of a covariate or treatment on expenditures, the MTP model provided less biased estimates with appropriate coverage probabilities for the effect, even when the distributional assumptions of the model were not met. The GLMs consistently provided higher bias and lower coverage for the effects of covariates than did the MTP model, particularly with a higher percentage of zeros. In particular, the estimated treatment effects from the GLMs were negatively biased, often suggesting that the treatment had the opposite effect on expenditures than it actually did.

Additionally, an analyst may want to consider the implications of a type I error. Even with only 20% zeros, the GLMs with empirical sandwich standard errors incurred inflated

type I error rates, and with 40% zeros, the GLMs incurred type I errors in one-quarter to one-half of all cases. If the cost of a type I error were high, the MTP model may be a safer alternative, with type I error rates never reaching higher than 7%.

On the other hand, if an analyst's primary goal is prediction of expenditures, the MTP model may not be a preferred option if the distributional assumptions are in question. The MTP model provided biased results with low coverage probabilities for total cost prediction when data were generated from the GG distribution, and these results did not appear to improve with increasing sample size. The GLMs, however, provided reasonably low bias in total cost prediction when the dataset had 20% zeros and the sample size was large.

If one were interested in prediction with a larger percentage of zeros when the distributional assumptions of the MTP were not met, none of the models we examined provided good predictions in this scenario. Particularly with larger sample sizes, the MTP model showed increased bias and lower coverage for predictions when data arose from the GG distribution. While decreasing with sample size in the GLMs, bias was still substantial and coverage fell well below the nominal 0.95 level when the data contained a larger percentage of zeros, even with the largest sample sizes assessed. To accommodate such cases, when interest is in prediction with questionable distributional assumptions and a substantial proportion of zeros, future work may be needed to find methods that accommodate a large proportion of zeros with less reliance on parametric assumptions. The MTP model could be extended to fit the GG distribution, and addition of empirical standard errors to the MTP model may increase coverage probabilities for predictions under questionable distributional assumptions. Regardless of modeling approach chosen, however, analysts will continue to need to carefully balance trade-offs in model fit, robustness, and interpretability with their specific analytic goals in mind.

Table 4.1: Descriptive statistics on LSN simulated data

	20% zeros			*40% zeros*		
Sample Size	Percent Zeros	Mean (SD)	Median (Q1-Q3)	Percent Zeros	Mean (SD)	Median (Q1-Q3)
200	21%	466 (1511)	184 (41-478)	40%	472 (9082)	89 (0-351)
1,000	21%	470 (1804)	184 (40-476)	40%	452 (6489)	88 (0-351)
10,000	21%	472 (2368)	184 (41-477)	40%	466 (12301)	88 (0-351)

Table 4.2: Median bias of estimated regression coefficients and total cost predictions in the marginal mean model from LSN data

				20% zeros				*40% zeros*		
Parameter	True Value	n	MTP	GLM $\lambda = 0$	GLM $\lambda = 1$	GLM $\lambda = 2$	MTP	GLM $\lambda = 0$	GLM $\lambda = 1$	GLM $\lambda = 2$
β_0	6	200	-0.02	-0.08	-0.07	-0.08	-0.02	-0.19	-0.18	-0.15
		1,000	-0.0009	-0.04	-0.03	-0.03	-0.002	-0.15	-0.14	-0.11
		10,000	0.0003	-0.04	-0.006	-0.005	0.0009	-0.07	-0.06	-0.05
β_1	0.2	200	0.0009	-0.05	-0.08	-0.13	0.008	-0.29	-0.43	-0.91
		1,000	0.005	-0.01	-0.03	-0.05	0.01	-0.16	-0.26	-0.47
		10,000	-0.0004	-0.01	-0.01	-0.01	-0.001	-0.09	-0.12	-0.17
β_2	-0.01	200	-0.002	0.06	0.10	0.16	-0.001	0.25	0.35	0.79
		1,000	0.0001	0.06	0.07	0.08	-0.002	0.21	0.28	0.47
		10,000	0.0004	0.06	0.03	0.03	-0.0003	0.14	0.16	0.20
β_3	0.05	200	-0.003	0.01	0.03	0.06	-0.002	0.05	0.09	0.24
		1,000	0.003	0.02	0.02	0.03	0.003	0.06	0.09	0.16
		10,000	0.00004	0.02	0.007	0.008	0.0001	0.04	0.05	0.06
Total Cost		200	-9.47	-50.19	-35.52	-30.44	-9.20	-112.26	-95.16	-82.25
		1,000	-0.81	-19.86	-14.11	-11.26	0.36	-64.72	-58.91	-48.93
		10,000	-0.26	-3.86	-2.48	-1.79	0.42	-26.07	-20.02	-14.49

Table 4.3: Coverage of 95% Wald-type confidence intervals for the marginal mean model parameters and total costs predictions from LSN data

		20% zeros				*40% zeros*			
Parameter	n	MTP	GLM $\lambda = 0$	GLM $\lambda = 1$	GLM $\lambda = 2$	MTP	GLM $\lambda = 0$	GLM $\lambda = 1$	GLM $\lambda = 2$
β_0	200	0.944	0.881	0.866	0.850	0.936	0.806	0.788	0.833
	1,000	0.953	0.883	0.881	0.882	0.951	0.764	0.740	0.817
	10,000	0.959	0.882	0.921	0.926	0.958	0.737	0.760	0.838
β_1	200	0.936	0.856	0.883	0.842	0.939	0.684	0.673	0.482
	1,000	0.932	0.904	0.903	0.880	0.956	0.746	0.695	0.552
	10,000	0.960	0.904	0.929	0.916	0.954	0.751	0.726	0.664
β_2	200	0.945	0.877	0.834	0.739	0.944	0.682	0.555	0.334
	1,000	0.948	0.821	0.817	0.756	0.956	0.553	0.472	0.351
	10,000	0.949	0.821	0.830	0.820	0.948	0.557	0.544	0.485
β_3	200	0.941	0.884	0.899	0.810	0.934	0.805	0.790	0.566
	1,000	0.960	0.906	0.888	0848	0.952	0.776	0.720	0.579
	10,000	0.946	0.906	0.909	0.892	0.947	0.751	0.728	0.683
Total Cost	200	0.924	0.855	0.857	0.815	0.907	0.704	0.696	0.583
	1,000	0.949	0.873	0.875	0.854	0.951	0.718	0.683	0.611
	10,000	0.956	0.898	0.903	0.896	0.952	0.723	0.714	0.687

Table 4.4: Descriptive statistics on data simulated from the generalized gamma distribution

	20% zeros			*40% zeros*		
Sample Size	Percent Zeros	Mean (SD)	Median (Q1-Q3)	Percent Zeros	Mean (SD)	Median (Q1-Q3)
200	21%	475 (1703)	188 (23-538)	40%	498 (15784)	74 (0-390)
1,000	21%	471 (1664)	186 (22-535)	40%	470 (13132)	72 (0-387)
10,000	21%	473 (6344)	187 (22-537)	40%	460 (7907)	72 (0-389)

Table 4.5: Median bias of estimated regression coefficients and total cost predictions in the marginal mean model from GG data

			20% zeros				*40% zeros*			
Parameter	True Value	n	MTP	GLM $\lambda = 0$	GLM $\lambda = 1$	GLM $\lambda = 2$	MTP	GLM $\lambda = 0$	GLM $\lambda = 1$	GLM $\lambda = 2$
β_0	6	200	0.02	-0.08	-0.05	-0.06	0.02	-0.17	-0.16	-0.12
		1,000	0.02	-0.04	-0.03	-0.03	-0.003	-0.13	-0.12	-0.10
		10,000	0.17	-0.009	-0.006	-0.005	0.04	-0.07	-0.06	-0.05
β_1	0.2	200	-0.008	-0.03	-0.07	-0.11	0.004	-0.24	-0.42	-0.89
		1,000	0.005	-0.02	-0.03	-0.04	0.002	-0.18	-0.27	-0.46
		10,000	-0.002	-0.01	-0.01	-0.02	-0.003	-0.10	-0.12	-0.17
β_2	-0.01	200	-0.003	0.07	0.09	0.14	0.009	0.23	0.36	0.83
		1,000	0.002	0.06	0.06	0.08	-0.003	0.21	0.28	0.47
		10,000	0.001	0.03	0.03	0.03	<0.0001	0.14	0.17	0.21
β_3	0.05	200	-0.003	0.01	0.03	0.05	-0.003	0.06	0.09	0.25
		1,000	0.001	0.02	0.02	0.03	0.0009	0.06	0.08	0.15
		10,000	0.0009	0.006	0.007	0.008	0.001	0.04	0.05	0.06
Total Cost		200	3.18	-35.57	-24.04	-19.35	6.94	-101.05	-86.45	-74.79
		1,000	11.44	-14.14	-9.62	-7.36	-1.99	-60.39	-51.04	-38.35
		10,000	79.86	-3.55	-2.36	-1.67	27.14	-27.30	-22.23	-16.72

Table 4.6: Coverage of 95% Wald-type confidence intervals for the marginal mean model parameters and total costs predictions from GG data

		20% zeros				*40% zeros*			
Parameter	n	MTP	GLM $\lambda = 0$	GLM $\lambda = 1$	GLM $\lambda = 2$	MTP	GLM $\lambda = 0$	GLM $\lambda = 1$	GLM $\lambda = 2$
β_0	200	0.946	0.917	0.905	0.890	0.938	0.828	0.823	0.848
	1,000	0.863	0.906	0.912	0.915	0.943	0.756	0.734	0.825
	10,000	0.348	0.903	0.905	0.916	0.509	0.703	0.725	0.795
β_1	200	0.944	0.870	0.887	0.851	0.944	0.710	0.686	0.475
	1,000	0.956	0.905	0.916	0.896	0.948	0.715	0.667	0.538
	10,000	0.955	0.907	0.899	0.889	0.938	0.738	0.719	0.664
β_2	200	0.940	0.864	0.824	0.735	0.940	0.669	0.545	0.322
	1,000	0.944	0.808	0.797	0.756	0.941	0.512	0.448	0.350
	10,000	0.960	0.806	0.802	0.793	0.951	0.527	0.507	0.464
β_3	200	0.928	0.877	0.880	0.797	0.924	0.792	0.777	0.560
	1,000	0.945	0.901	0.906	0.863	0.945	0.750	0.705	0.596
	10,000	0.952	0.889	0.882	0.871	0.936	0.724	0.693	0.651
Total Cost	200	0.929	0.873	0.876	0.834	0.926	0.720	0.708	0.581
	1,000	0.893	0.884	0.888	0.869	0.937	0.697	0.670	0.610
	10,000	0.353	0.886	0.887	0.882	0.539	0.702	0.693	0.666

Table 4.7: Type I error rates at nominal significance level 0.05 for LSN and GG data

		20% zeros				*40% zeros*		
			GLM	GLM	GLM		GLM	GLM	GLM
Distribution	n	MTP	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$	MTP	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$
LSN	200	0.068	0.157	0.115	0.158	0.062	0.328	0.325	0.518
	1,000	0.066	0.100	0.098	0.120	0.044	0.265	0.305	0.448
	10,000	0.042	0.073	0.071	0.084	0.047	0.256	0.274	0.335
GG	200	0.054	0.133	0.113	0.149	0.057	0.304	0.316	0.524
	1,000	0.047	0.099	0.084	0.104	0.051	0.304	0.333	0.462
	10,000	0.044	0.097	0.101	0.111	0.055	0.264	0.280	0.336

CHAPTER 5: CONCLUSION

Analyzing semicontinuous data, such as medical expenditures, has posed challenges to analysts for decades. Modeling approaches must appropriately account for the unique statistical properties of semicontinuous data, but at the same time, investigators need model estimates that are interpretable for their policy questions of interest. Previously, methods have not existed that simultaneously accounted for the excess zeros and skewness while also providing easily interpretable estimates of covariate effects on the overall marginal mean, $\mathrm{E}(Y)$.

This dissertation developed a new marginalized two-part (MTP) model that overcomes many of the drawbacks of previous approaches, including difficulty in interpreting covariate effects on the overall mean, a target of primary interest in many studies. Rather than parameterizing the model in terms of the mean of the transformed, conditionally positive outcomes in the second part, the MTP model parameterized covariate effects directly on the overall mean, $\mathrm{E}(Y)$, on the untransformed scale. This allows parameter estimates to be interpreted as the multiplicative effect on the overall mean rather than on the conditional mean of only the positive outcomes. Our approach also has the advantage of providing estimates of covariate effects on the probability of incurring a positive-valued outcome, as in the first part of two-part models, as well as accounting for the zero-inflated and skewed nature of many semicontinuous outcomes.

We extended the MTP model to longitudinal data via the inclusion of random effects. This model could be fit using maximum likelihood or Bayesian approaches, although we proposed the latter to increase flexibility to model complex random effect structures. Specifically, we fit correlated random effects to allow dependence between the probability of incurring a positive outcome and the level of the outcome. This approach provided easily computed predictions of the overall mean outcome, and the parameter estimates obtained from the MTP model

provided the same simple interpretation as those from the one-part GLMs without sacrificing statistical appropriateness. Additionally, while the model was subject-specific with random effects, many parameters had dual interpretations as both subject-specific and population average, further increasing interpretability of model results. Thus, the MTP model can provide useful policy conclusions while remaining rooted in good statistical practice.

Finally, we compared one-part GLMs fit using quasi-likelihood with the MTP model under a variety of simulated data generating mechanisms to assess under which scenarios one may be able to fit the simpler one-part models without inducing too much bias or sacrificing too much precision, or alternatively, when two-part models are needed for appropriate statistical inference. One-part models, while simpler to fit, often displayed substantial bias, particularly when the percentage of zeros was higher. Although it decreased with sample size, we found that even with a sample size of 10,000, bias was still problematic. Similarly, under-coverage of nominal 95% confidence intervals for parameters and model predictions was also problematic for the GLMs in the presence of many zero-valued observations. The MTP model, on the other hand, provided very low bias and appropriate coverage of covariate effects, even when the distributional assumptions were not met. However, when the parametric assumptions for the MTP were not met, bias increased and coverage dropped for model predictions.

The MTP models provide a step forward in the quest for statistically appropriate and interpretable analytic methods for semicontinuous data. Future research is needed to extend these methods in many other directions, such as the addition of more flexible distributions, the development of model evaluation approaches, or the incorporation of methods for spatially correlated data.

APPENDIX A: SAS CODE FROM CHAPTER 2

Marginalized log-normal model

```
proc nlmixed data=mydata;
    bounds 0 <= sigma2;
    parms /* initial values for parameters */ ;
    linbin = a0 + a1*x1 + a2*x2;
    binprob = exp(linbin)/(1+exp(linbin)); /* probability y > 0 */
    mu = b0 + b1*x1 + b2*x2 - log(binprob) - sigma2/2;
    if y=0 then loglik=log(1-binprob);
    else if y>0 then loglik=log(binprob)-log(y)-.5*log(2*CONSTANT('PI'))
            -log(sqrt(sigma2))-(1/(2*sigma2))*(log(y)-mu)**2;
    model y~general(loglik);
    estimate 'marginal mean at x1=50 and x2=1' exp(b0+b1*50+b2*1);
run;
```

Marginalized log-skew-normal model

```
proc nlmixed data=mydata;
    bounds 0<=omega;
    parms  /* initial values for parameters */ ;
    linbin = a0 + a1*x1 + a2*x2;
    binprob = exp(linbin)/(1+exp(linbin)); /* probability y > 0 */
    delta = kappa/sqrt(1+kappa**2);
    xi = b0 + b1*x1 + b2*x2 - log(2) - log(binprob)
           - log(CDF('NORMAL',omega*delta, 0, 1))-(omega**2)/2;
    if y=0 then loglik=log(1-binprob);
    else if y>0 then do;
        pdfnormvar=(log(y)-xi)/omega;
        cdfnormvar=kappa*((log(y)-xi)/omega);
        loglik=log(binprob)+log(2)-log(y)-log(omega)
```

```
                +log(PDF('NORMAL', pdfnormvar, 0, 1))
                +log(CDF('NORMAL', cdfnormvar, 0,1));
    end;
    model y~general(loglik);
    estimate 'marginal mean at x1=50 and x2=1' exp(b0+b1*50+b2*1);
run;
```

APPENDIX B: DERIVATION OF $\mathrm{E}(Y_{IJ})$ FROM CHAPTER 3

Recall that the random effects, $\boldsymbol{b}_i' = (\boldsymbol{a}_i', \boldsymbol{d}_i')$, are specified as in equation (3.4). We must find:

$$\begin{aligned}
\mathrm{E}(Y_{ij}) &= \mathrm{E}_{\boldsymbol{b}_i}\left\{\mathrm{E}\left[Y_{ij}|\boldsymbol{b}_i\right]\right\} = \mathrm{E}_{\boldsymbol{b}_i}\left[\mathrm{e}^{\boldsymbol{x}_{2ij}'\boldsymbol{\beta}+\boldsymbol{z}_{2ij}'\boldsymbol{d}_i}\right] \\
&= \mathrm{E}_{\boldsymbol{b}_i}\left[\mathrm{e}^{\boldsymbol{x}_{2ij}'\boldsymbol{\beta}}\right]\cdot \mathrm{E}_{\boldsymbol{b}_i}\left[\mathrm{e}^{\boldsymbol{z}_{2ij}'\boldsymbol{d}_i}\right] \\
&= \mathrm{e}^{\boldsymbol{x}_{2ij}'\boldsymbol{\beta}}\cdot \mathrm{E}_{\boldsymbol{b}_i}\left[\mathrm{e}^{\boldsymbol{z}_{2ij}'\boldsymbol{d}_i}\right]. \qquad \text{(B.1)}
\end{aligned}$$

Now consider $\mathrm{E}_{\boldsymbol{b}_i}\left[\mathrm{e}^{\boldsymbol{z}_{2ij}'\boldsymbol{d}_i}\right] = \mathrm{E}_{\boldsymbol{a}_i}\left\{\mathrm{E}_{\boldsymbol{d}_i}\left[\mathrm{e}^{\boldsymbol{z}_{2ij}'\boldsymbol{d}_i}|\boldsymbol{a}_i\right]\right\}$. Assuming $\boldsymbol{b}_i$ follows the multivariate normal distribution as shown in equation (3.4), we have $\boldsymbol{d}_i|\boldsymbol{a}_i \sim N(\boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{a}_i, \boldsymbol{\Sigma}_{dd} - \boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad})$. Utilizing the moment generating function of the multivariate normal distribution, $M_{\boldsymbol{d}_i}(t) = \mathrm{E}\left(\mathrm{e}^{t\boldsymbol{d}_i}\right)$ with $t = 1$, we therefore have

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{b}_i}\left[\mathrm{e}^{\boldsymbol{z}_{2ij}'\boldsymbol{d}_i}\right] &= \mathrm{E}_{\boldsymbol{a}_i}\left\{\mathrm{E}_{\boldsymbol{d}_i}\left[\mathrm{e}^{\boldsymbol{z}_{2ij}'\boldsymbol{d}_i}|\boldsymbol{a}_i\right]\right\} \\
&= \mathrm{E}_{\boldsymbol{a}_i}\left[\exp\left(\boldsymbol{z}_{2ij}'\boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{a}_i + \frac{1}{2}\boldsymbol{z}_{2ij}'\left(\boldsymbol{\Sigma}_{dd} - \boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad}\right)\boldsymbol{z}_{2ij}\right)\right] \\
&= \mathrm{E}_{\boldsymbol{a}_i}\left[\exp\left(\boldsymbol{z}_{2ij}'\boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{a}_i\right)\right]\cdot \exp\left[\frac{1}{2}\boldsymbol{z}_{2ij}'\left(\boldsymbol{\Sigma}_{dd} - \boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad}\right)\boldsymbol{z}_{2ij}\right]. \qquad \text{(B.2)}
\end{aligned}$$

Now, because $\boldsymbol{a}_i$ is marginally distributed as $N(\boldsymbol{0}, \boldsymbol{\Sigma}_{aa})$, it follows that

$$\begin{aligned}
\boldsymbol{z}_{2ij}'\boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{a}_i &\sim N\left(0, \boldsymbol{z}_{2ij}'\boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{aa}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad}\boldsymbol{z}_{2ij}\right) \\
&\sim N\left(0, \boldsymbol{z}_{2ij}'\boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad}\boldsymbol{z}_{2ij}\right).
\end{aligned}$$

Then using the moment-generating function of the univariate normal distribution, we find

$$\mathrm{E}_{\boldsymbol{a}_i}\left[\exp\left(\boldsymbol{z}_{2ij}'\boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{a}_i\right)\right] = \exp\left[\frac{1}{2}\boldsymbol{z}_{2ij}'\boldsymbol{\Sigma}_{ad}'\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad}\boldsymbol{z}_{2ij}\right].$$

Plugging this back into (B.2), we have

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{b}_i}\left[\mathrm{e}^{\boldsymbol{z}'_{2ij}\boldsymbol{d}_i}\right] &= \exp\left[\frac{1}{2}\boldsymbol{z}'_{2ij}\boldsymbol{\Sigma}'_{ad}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad}\boldsymbol{z}_{2ij} + \frac{1}{2}\boldsymbol{z}'_{2ij}\left(\boldsymbol{\Sigma}_{dd} - \boldsymbol{\Sigma}'_{ad}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad}\right)\boldsymbol{z}_{2ij}\right] \\
&= \exp\left[\frac{1}{2}\boldsymbol{z}'_{2ij}\left(\boldsymbol{\Sigma}'_{ad}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad} + \boldsymbol{\Sigma}_{dd} - \boldsymbol{\Sigma}'_{ad}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ad}\right)\boldsymbol{z}_{2ij}\right] \\
&= \exp\left(\frac{1}{2}\boldsymbol{z}'_{2ij}\boldsymbol{\Sigma}_{dd}\boldsymbol{z}_{2ij}\right).
\end{aligned}$$

Plugging this back into equation (B.1) for $\mathrm{E}(Y_{ij})$, we have

$$\mathrm{E}(Y_{ij}) = \exp\left(\boldsymbol{x}'_{2ij}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}'_{2ij}\boldsymbol{\Sigma}_{dd}\boldsymbol{z}_{2ij}\right)$$

as the overall population average marginal mean.

Examining the effect of a unit increase in covariate x_{2kij} on this mean, where x_{2kij} is not an element of $\boldsymbol{z}_{2ij}$, we obtain

$$\begin{aligned}
\frac{\mathrm{E}(Y_{ij}|x_{2kij} = l+1, \boldsymbol{x}_{2(-k)ij})}{\mathrm{E}(Y_{ij}|x_{2kij} = l, \boldsymbol{x}_{2(-k)ij})} &= \frac{\exp\left(\boldsymbol{x}_{2(-k)ij}\boldsymbol{\beta}_{(-k)} + \beta_k \cdot (l+1) + \frac{1}{2}\left(\boldsymbol{z}'_{2ij}\boldsymbol{\Sigma}_{dd}\boldsymbol{z}_{2ij}\right)\right]}{\exp\left[\boldsymbol{x}_{2(-k)ij}\boldsymbol{\beta}_{(-k)} + \beta_k \cdot l + \frac{1}{2}\left(\boldsymbol{z}'_{2ij}\boldsymbol{\Sigma}_{dd}\boldsymbol{z}_{2ij}\right)\right]} \\
&= \exp(\beta_k).
\end{aligned}$$

APPENDIX C: SAS PROC MCMC CODE FROM CHAPTER 3

```
proc mcmc data=one nbi=10000 nmc=20000 thin=5 seed=41514
propcov=quanew dic statistics=all
monitor=(alph bet meancopay meanexempt copaymultiplicative copayadditive
kappa omegasq sigbi1-sigbi16);
array mu0a[8];
array sig0a[8,8];
array mu0b[8];
array sig0b[8,8];
array alph[8];
array bet[8];

/* Specify the random effects */
array sigbi[4,4];
array bi[4];
array mubi[4] (0 0 0 0 0);
array sig0bi[4,4] (1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1);

/* Arrays for desired functions of parameters */
array meancopay[4];
array meanexempt[4];
array copaymultiplicative[4];
array copayadditive[4];

begincnst;
call zeromatrix(mu0a); * prior mean of 0 for alphas;
call identity(Sig0a);
call mult(Sig0a, 1000, Sig0a);
call zeromatrix(mu0b); * prior mean of 0 for betas;
call identity(Sig0b);
call mult(Sig0b, 1000, Sig0b);
```

```
endcnst;

beginnodata;
delta=kappa/sqrt(1+kappa**2);
omega=sqrt(omegasq);

/* Model estimated population average overall means */
* year 2000, copay required;
meancopay[1]=exp(bet[1]+bet[5]+.5*(sigbi11));
* year 2001, copay required;
meancopay[2]=exp(bet[1]+bet[2]+bet[5]+bet[6]+.5*(sigbi11+2*sigbi12+sigbi16));
* year 2002, copay required;
meancopay[3]=exp(bet[1]+bet[3]+bet[5]+bet[7]+.5*(sigbi11+4*sigbi12+4*sigbi16));
* year 2003, copay required;
meancopay[4]=exp(bet[1]+bet[4]+bet[5]+bet[8]+.5*(sigbi11+6*sigbi12+9*sigbi16));
* year 2000, copay exempt;
meanexempt[1]=exp(bet[1]+.5*(sigbi11));
* year 2001, copay exempt;
meanexempt[2]=exp(bet[1]+bet[2]+.5*(sigbi11+2*sigbi12+sigbi16));
* year 2002, copay exempt;
meanexempt[3]=exp(bet[1]+bet[3]+.5*(sigbi11+4*sigbi12+4*sigbi16));
* year 2003, copay exempt;
meanexempt[4]=exp(bet[1]+bet[4]+.5*(sigbi11+6*sigbi12+9*sigbi16));

/* Multiplicative and additive effects of copay requirement */
* year 2000;
copaymultiplicative[1]=exp(bet[5]);
copayadditive[1]=exp(bet[1]+bet[5]+.5*(sigbi11))-exp(bet[1]+.5*(sigbi11));
* year 2001;
copaymultiplicative[2]=exp(bet[5]+bet[6]);
copayadditive[2]=exp(bet[1]+bet[2]+bet[5]+bet[6]+.5*(sigbi11+2*sigbi12+sigbi16))
-exp(bet[1]+bet[2]+.5*(sigbi11+2*sigbi12+sigbi16));
* year 2002;
```

```
copaymultiplicative[3]=exp(bet[5]+bet[7]);
copayadditive[3]=exp(bet[1]+bet[3]+bet[5]+bet[7]+.5*(sigbi11+4*sigbi12+4*sigbi16))
-exp(bet[1]+bet[3]+.5*(sigbi11+4*sigbi12+4*sigbi16));
* year 2003;
copaymultiplicative[4]=exp(bet[5]+bet[8]);
copayadditive[4]=exp(bet[1]+bet[4]+bet[5]+bet[8]+.5*(sigbi11+6*sigbi12+9*sigbi16))
-exp(bet[1]+bet[4]+.5*(sigbi11+6*sigbi12+9*sigbi16));
endnodata;

/* Specify parameters */
parm alph {0 0.3 0.6 0.3 -1.1 -.06 -.2 -.2};
parm bet {0 0.07 0.12 0.23 -.09 -.02 -.19 -.28};
parm kappa 1 omegasq 0.8;
parm sigbi;

/* Specify prior distributions */
prior alph ~ mvn(mu0a, sig0a);
prior bet ~ mvn(mu0b, sig0b);
prior sigbi ~ iwish(4, sig0bi); * random effects covariance prior;
prior kappa ~ uniform(-10,10);
prior omegasq ~ igamma(0.001, scale=0.001);

random bi ~ mvn(mean=mubi, cov=sigbi) subject=id;

linbin = alph[1] + alph[2]*yr2001 + alph[3]*yr2002 + alph[4]*yr2003
           + alph[5]*mustpay + alph[6]*payyr01 + alph[7]*payyr02 + alph[8]*payyr03
           + bi[1] + bi[2]*time;
binprob = exp(linbin)/(1+exp(linbin));
mu = bet[1] + bet[2]*yr2001 + bet[3]*yr2002 + bet[4]*yr2003 + bet[5]*mustpay
+ bet[6]*payyr01 + bet[7]*payyr02 + bet[8]*payyr03 + bi[3] + bi[4]*time
- log(2) - log(binprob) - log(CDF('NORMAL',omega*delta, 0, 1))-omega**2/2;

if spcost=0 then loglik=log(1-binprob);
```

```
else if spcost>0 then do;
pdfnormvar=(log(spcost)-mu)/omega;
cdfnormvar=kappa*((log(spcost)-mu)/omega);
loglik=log(binprob)+log(2)-log(spcost)
-log(omega)+log(PDF('NORMAL', pdfnormvar, 0, 1))
+  log(CDF('NORMAL', cdfnormvar, 0,1));
end;

model spcost~general(loglik);

run;
```

APPENDIX D: CONVERGENCE DIAGNOSTICS FROM CHAPTER 3

The final MTP model used in the analysis of the VA specialty care copayment increase in Section 3.5 was specified as:

$$\begin{aligned}
\text{logit}(\pi_{ij}) &= \alpha_0 + \alpha_1 \text{YR01}_{ij} + \alpha_2 \text{YR02}_{ij} + \alpha_3 \text{YR03}_{ij} + \alpha_4 \text{COPAY}_i + \alpha_5 \text{COPAY}_i \times \text{YR01}_{ij} \\
&\quad + \alpha_6 \text{COPAY}_i \times \text{YR02}_{ij} + \alpha_7 \text{COPAY}_i \times \text{YR03}_{ij} + a_{1i} + a_{2i} t_{ij}, \quad \text{and} \\
\text{E}(Y_{ij}|\boldsymbol{b}_i) &= \exp(\beta_0 + \beta_1 \text{YR01}_{ij} + \beta_2 \text{YR02}_{ij} + \beta_3 \text{YR03}_{ij} + \beta_4 \text{COPAY}_i + \beta_5 \text{COPAY}_i \times \text{YR01}_{ij} \\
&\quad + \beta_6 \text{COPAY}_i \times \text{YR02}_{ij} + \beta_7 \text{COPAY}_i \times \text{YR03}_{ij} + d_{1i} + d_{2i} t_{ij}),
\end{aligned}$$

where $t_{ij} = 0, 1, 2, 3$ for years 2000, 2001, 2002, and 2003, respectively. Convergence diagnostics for all model parameters are shown in the figures below. Due to indexing in SAS software, the diagnostics labeled "alph1" correspond to parameter α_0, those labeled "alph2" correspond to α_1, and so on.

Figure D.1: Convergence diagnostics for α_0

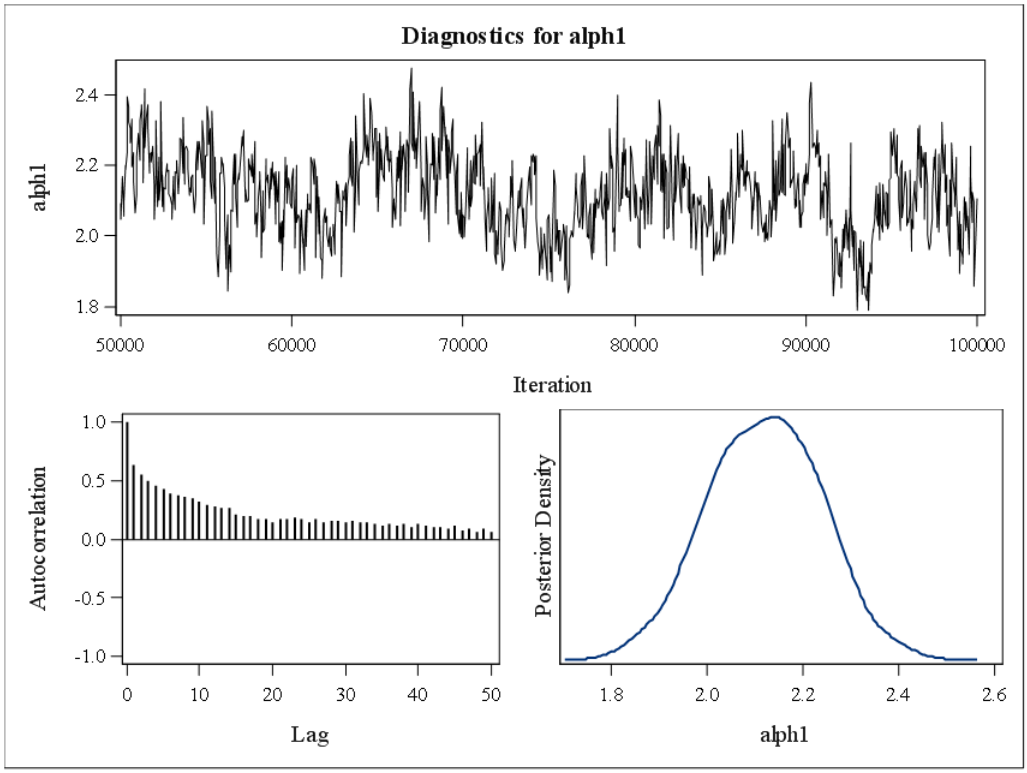


Figure D.2: Convergence diagnostics for α_1

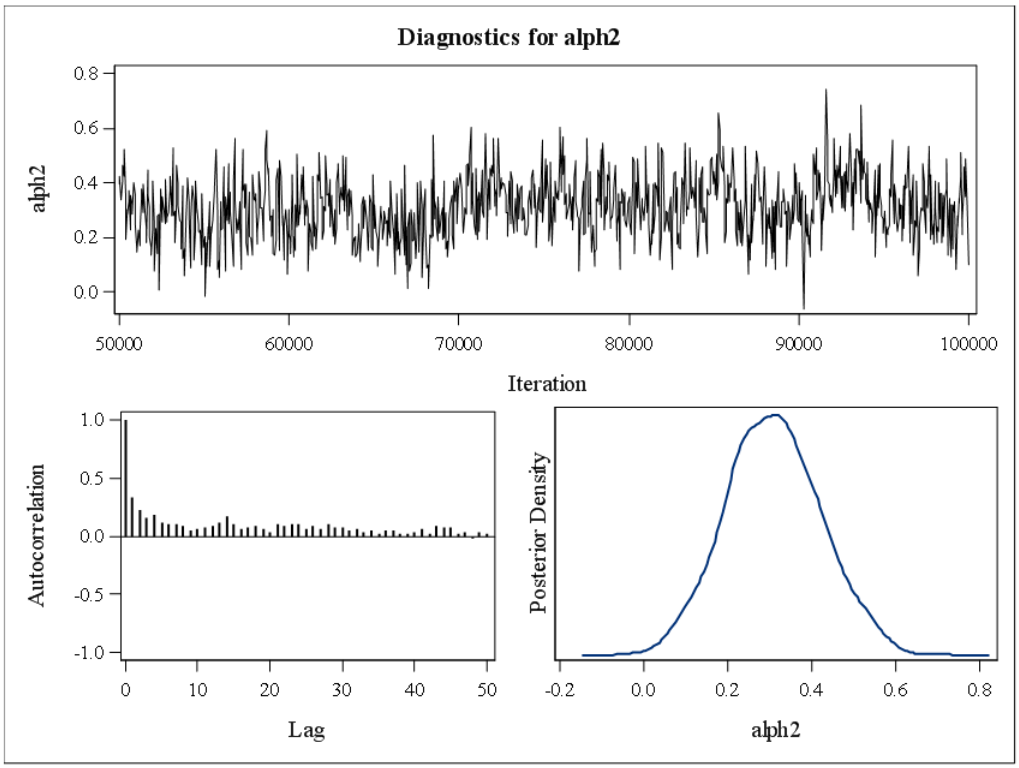


Figure D.3: Convergence diagnostics for α_2

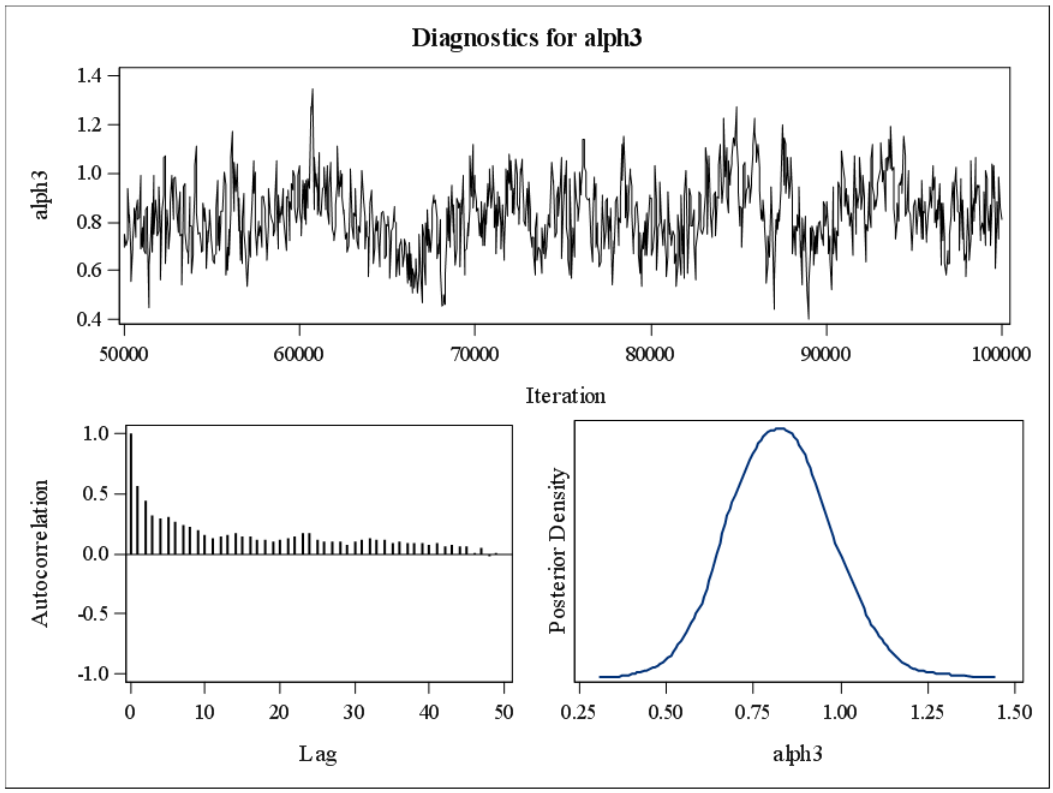


Figure D.4: Convergence diagnostics for α_3

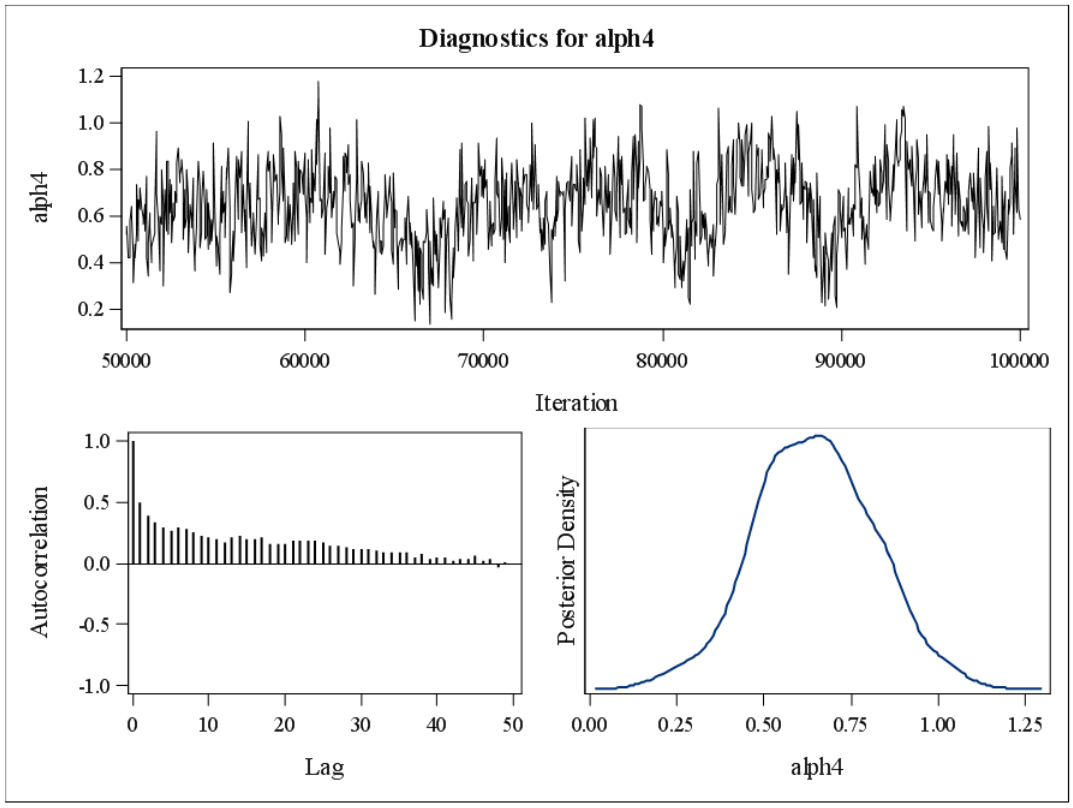


Figure D.5: Convergence diagnostics for α_4

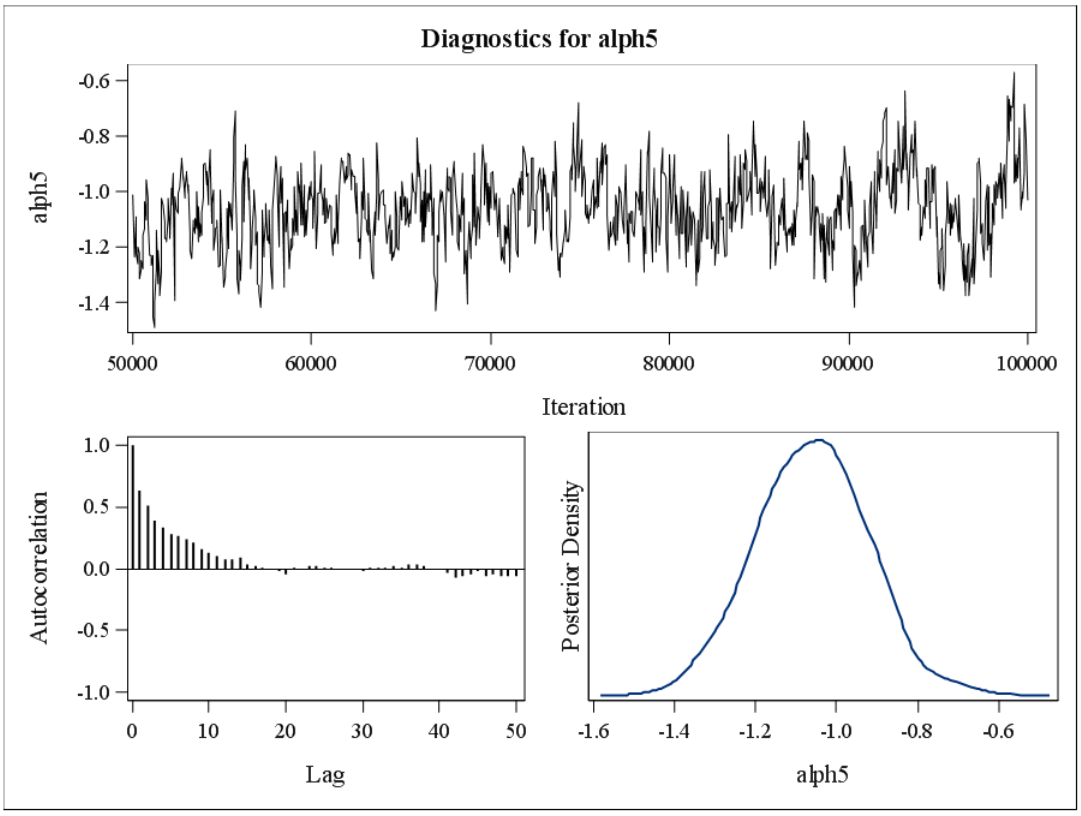


Figure D.6: Convergence diagnostics for α_5

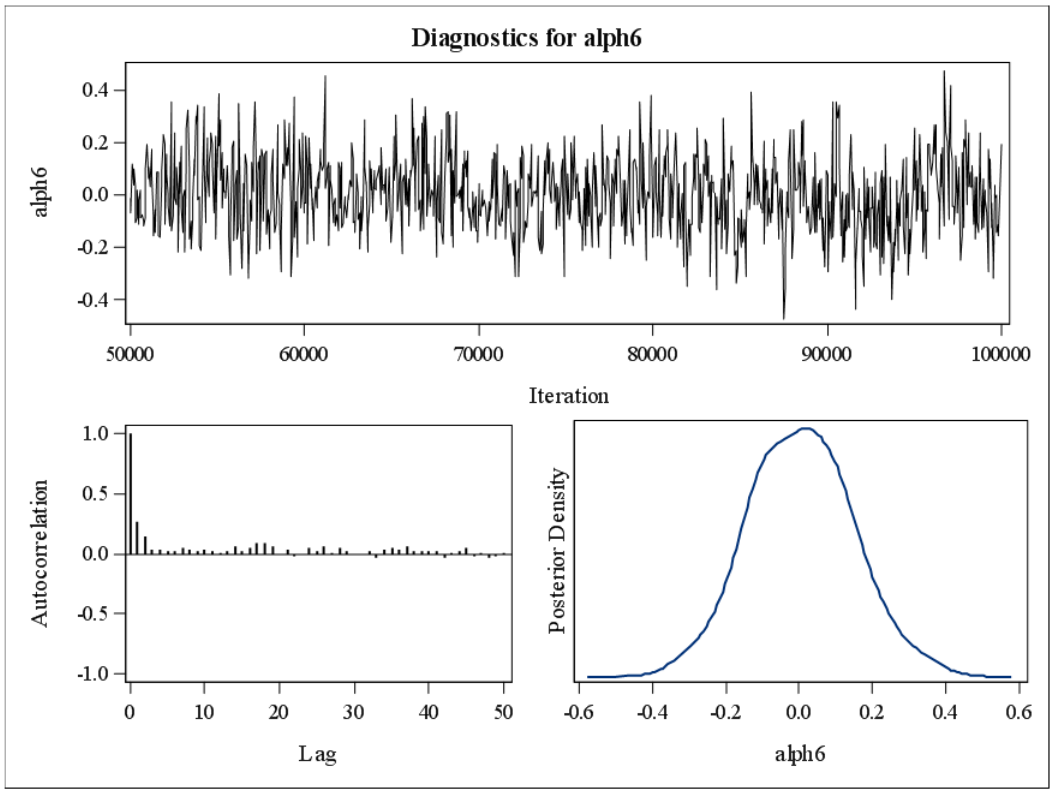


Figure D.7: Convergence diagnostics for α_6

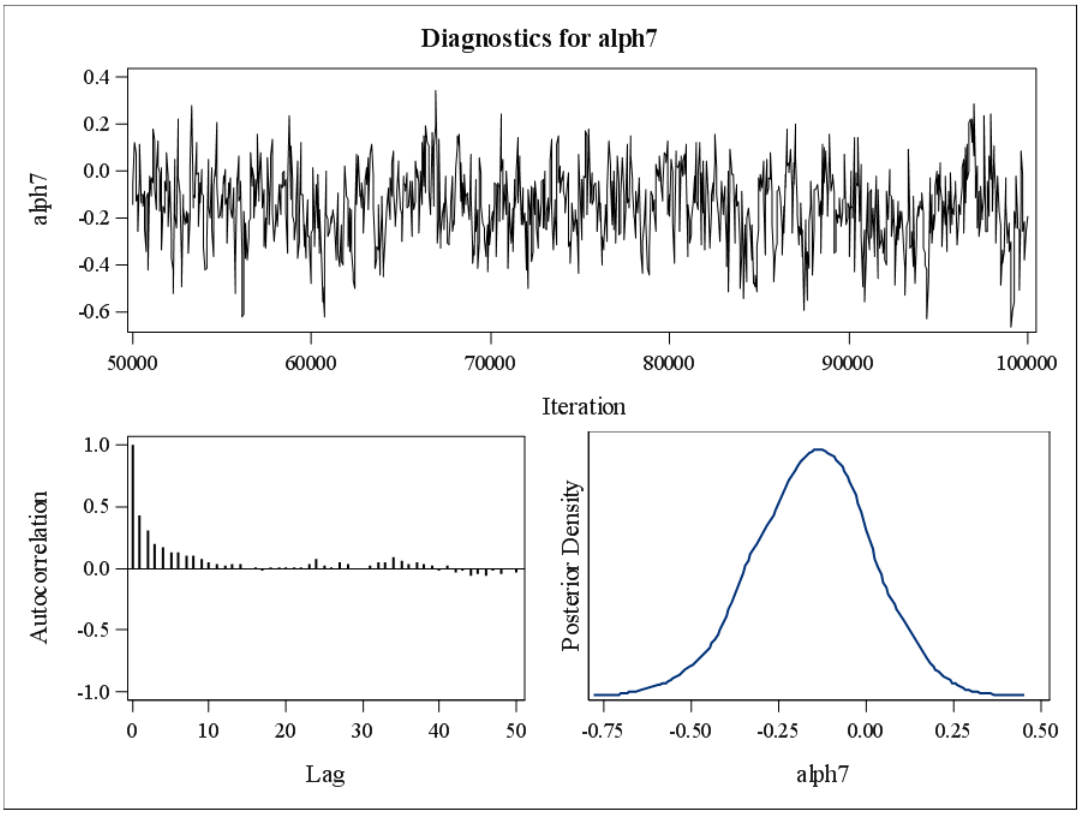


Figure D.8: Convergence diagnostics for α_7

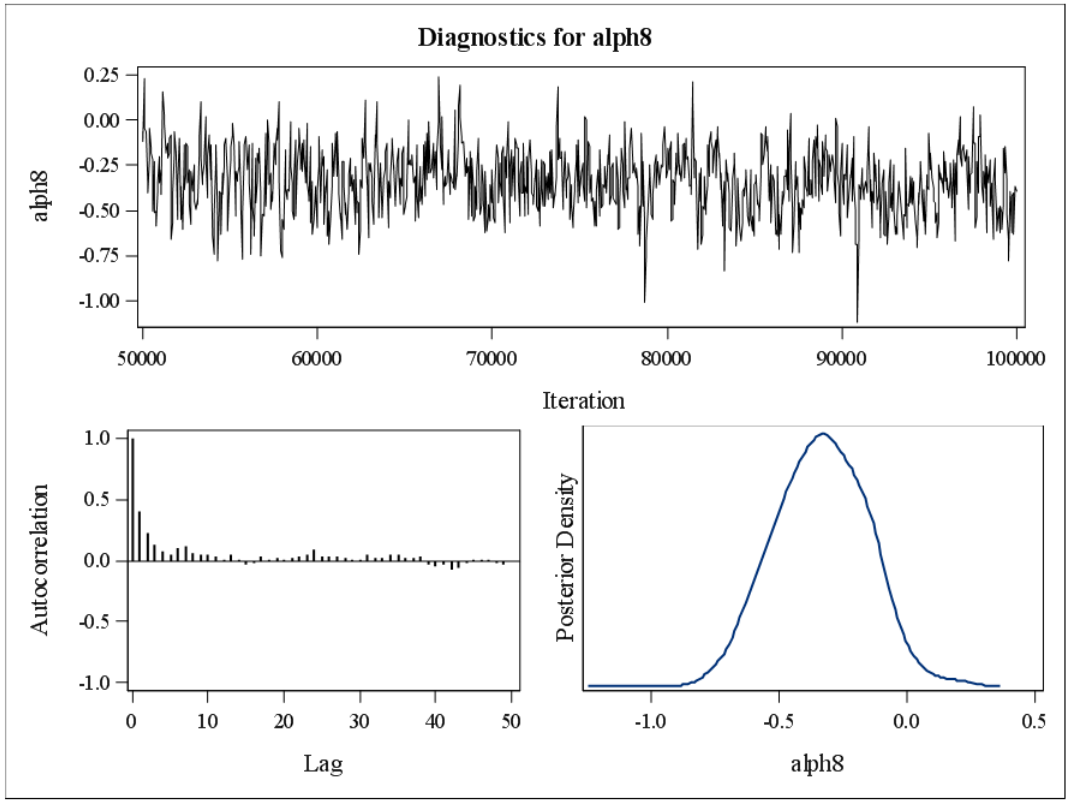


Figure D.9: Convergence diagnostics for β_0

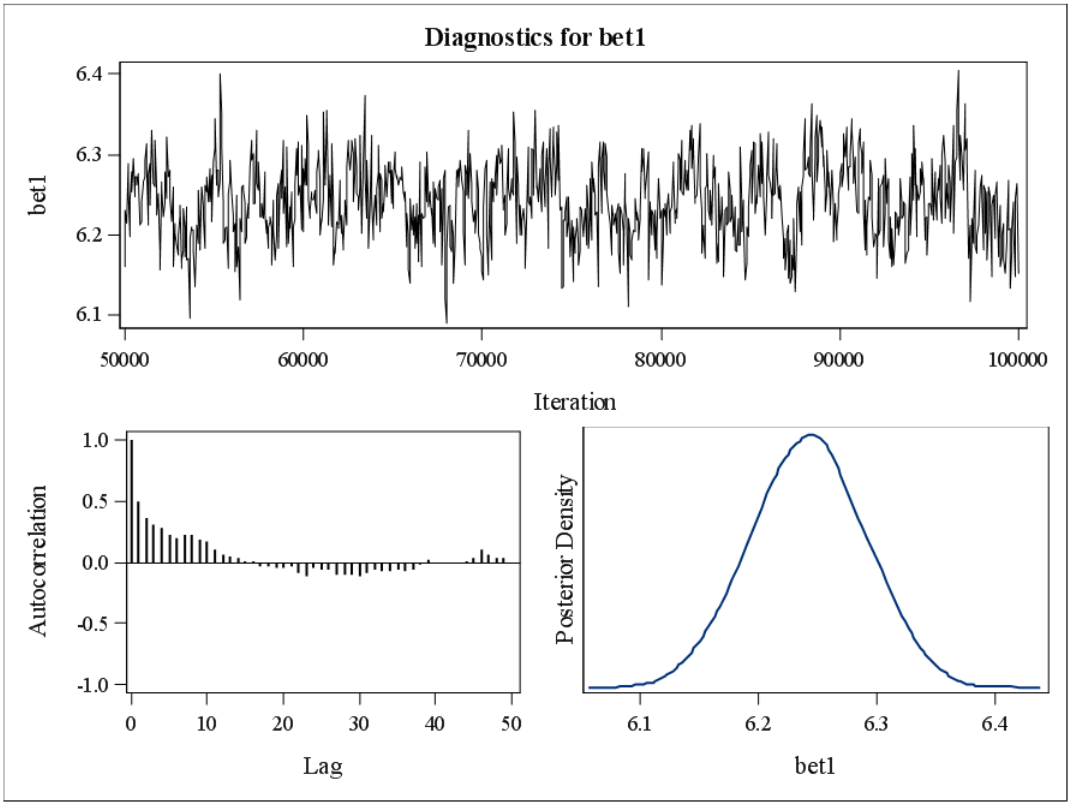


Figure D.10: Convergence diagnostics for β_1

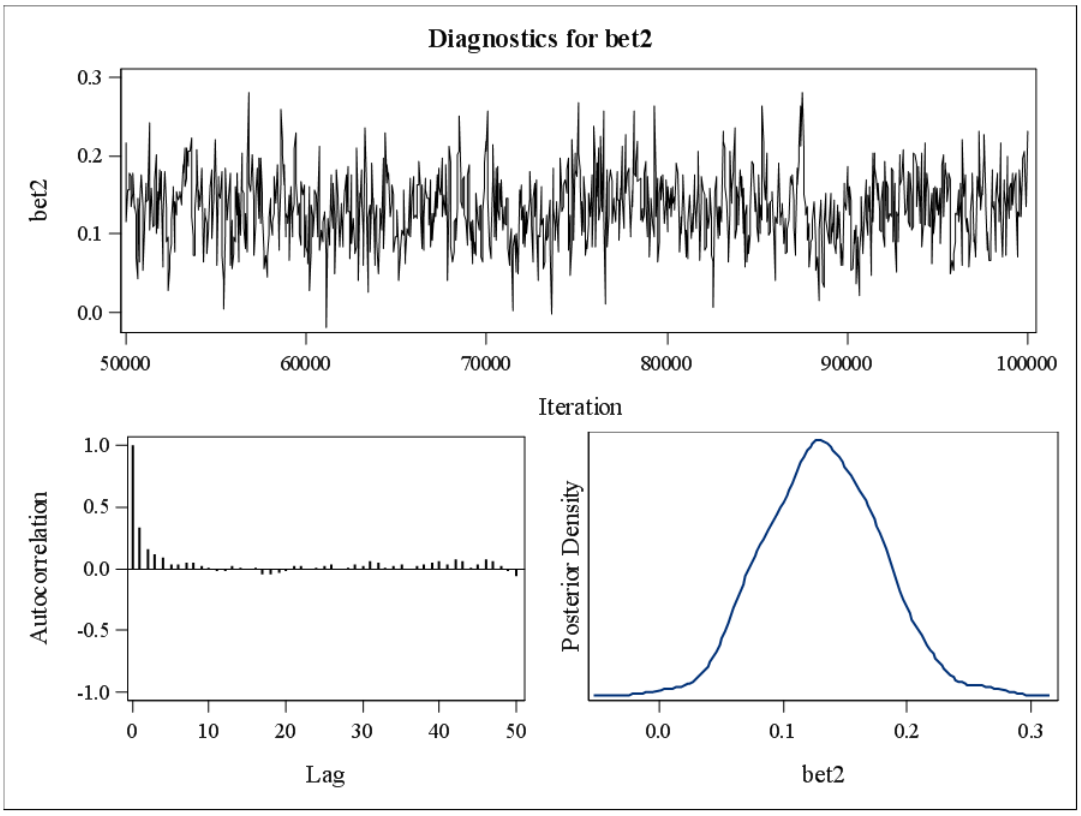


Figure D.11: Convergence diagnostics for β_2

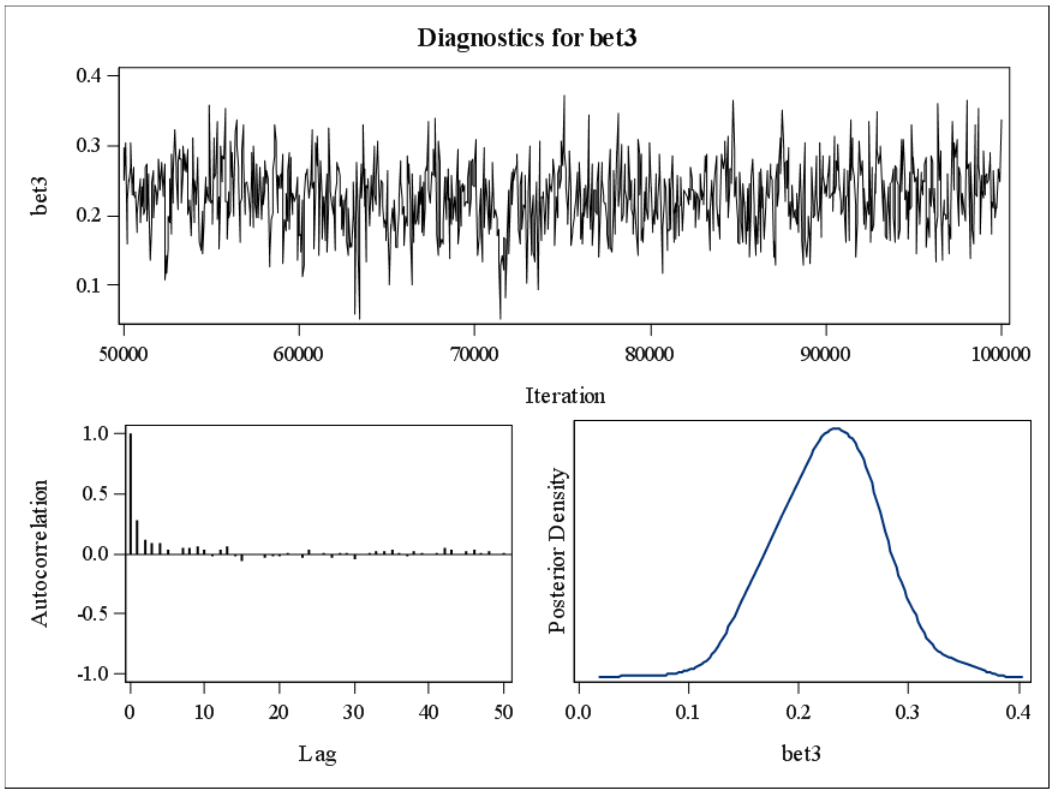


Figure D.12: Convergence diagnostics for β_3

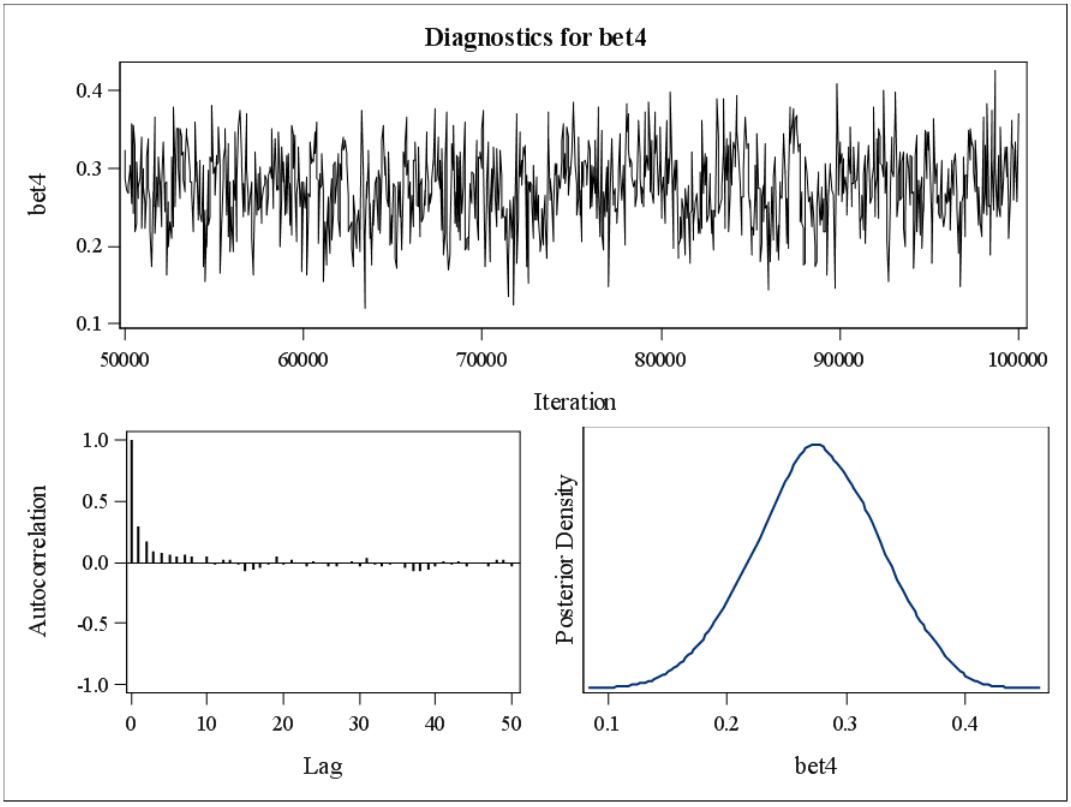


Figure D.13: Convergence diagnostics for β_4

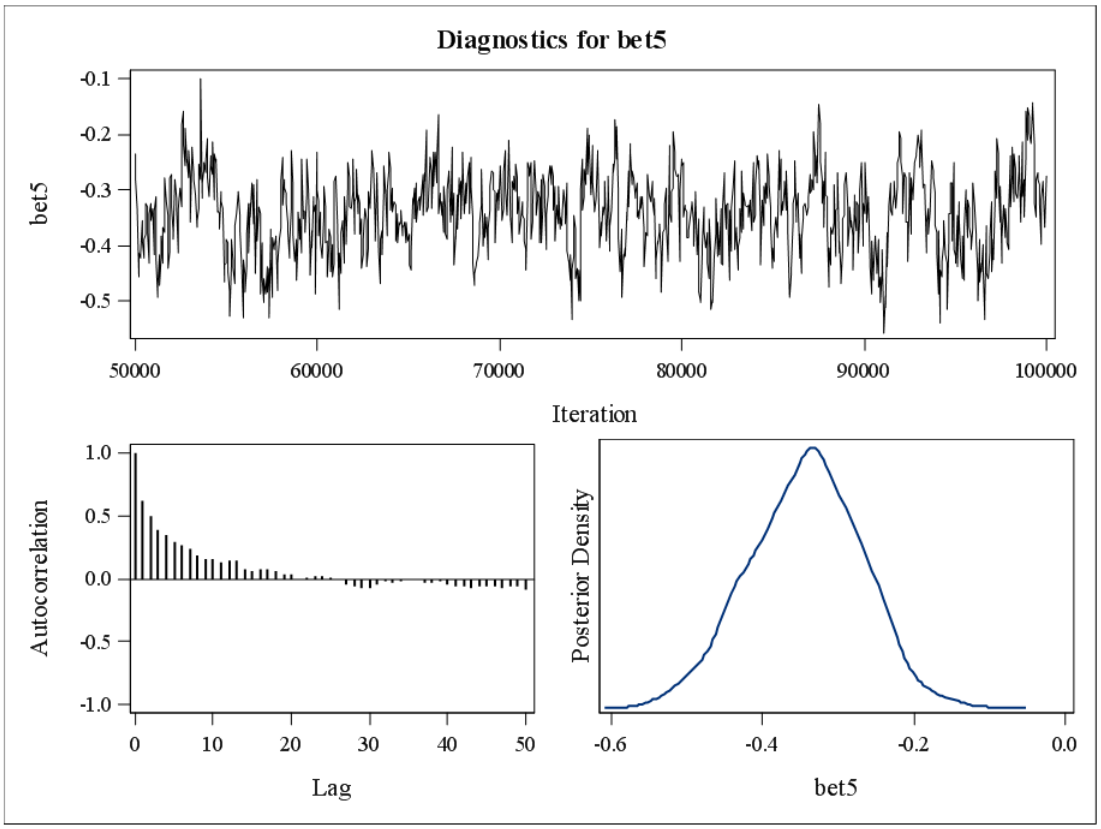


Figure D.14: Convergence diagnostics for β_5

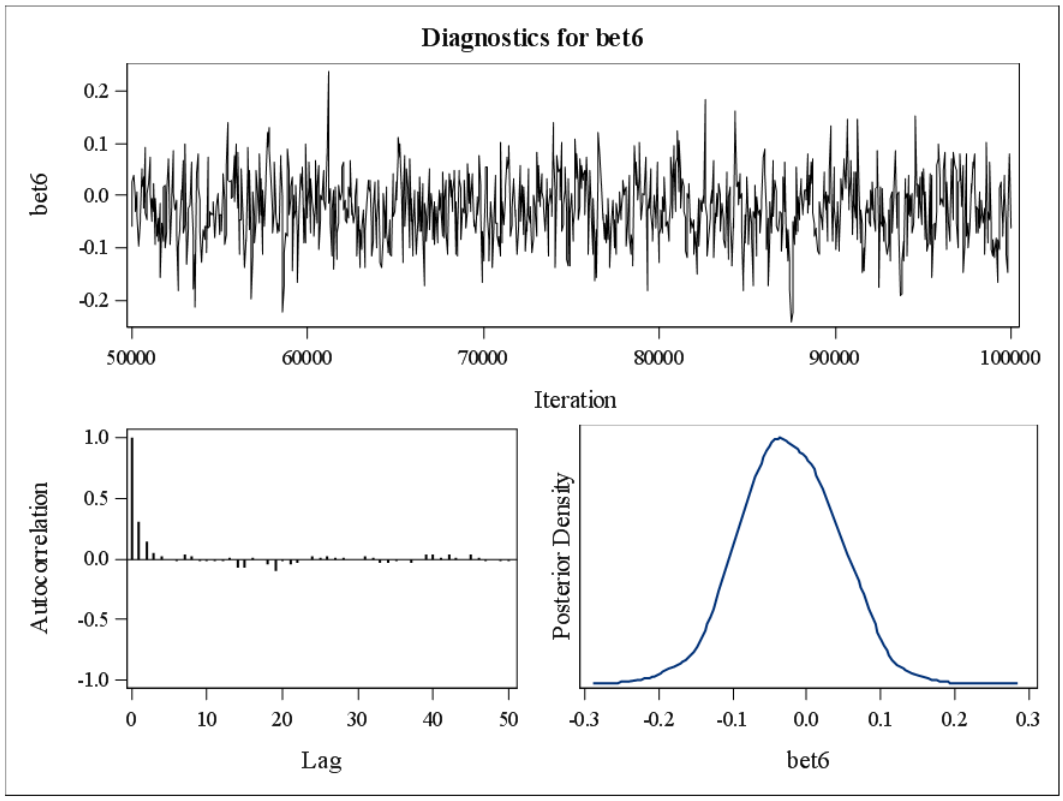


Figure D.15: Convergence diagnostics for β_6

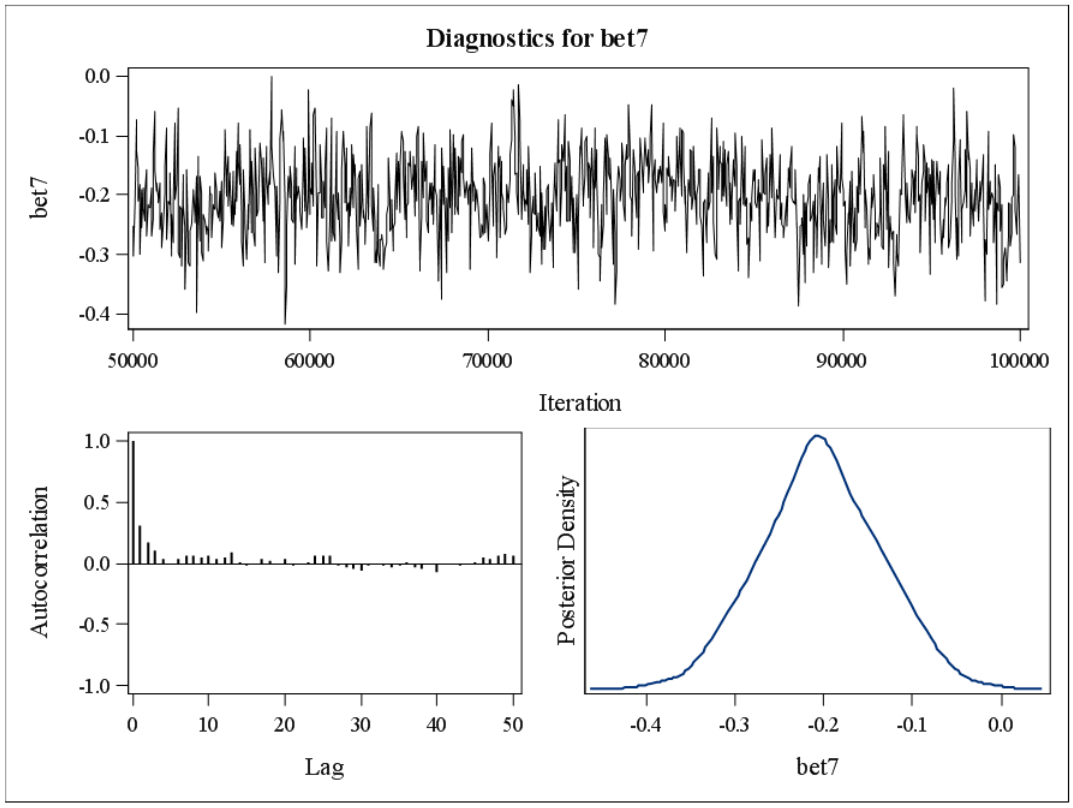


Figure D.16: Convergence diagnostics for β_7

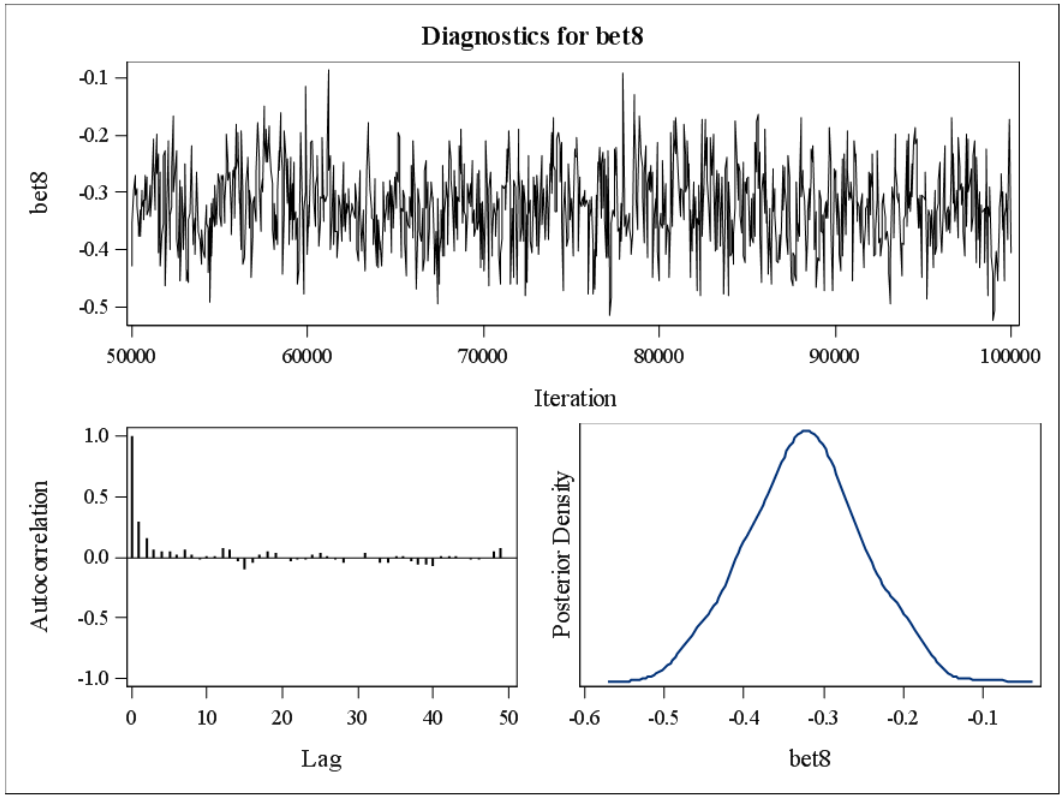


Figure D.17: Convergence diagnostics for scale parameter, ω^2

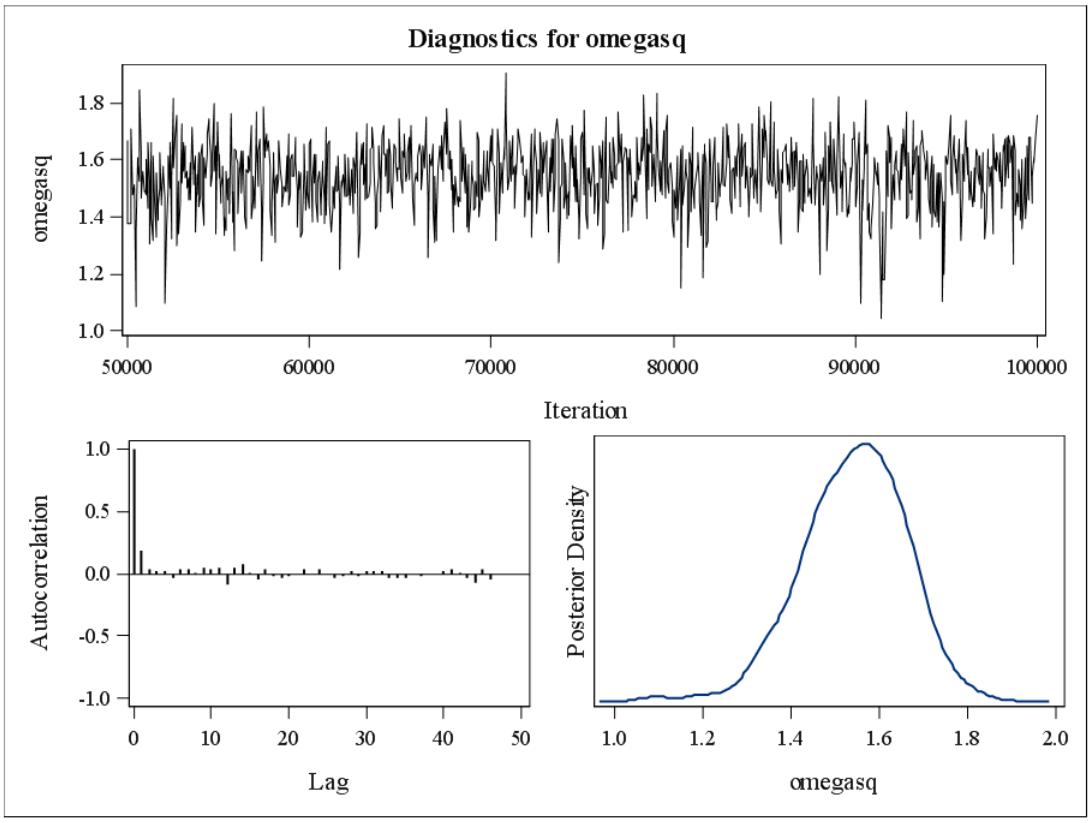


Figure D.18: Convergence diagnostics for shape parameter, κ

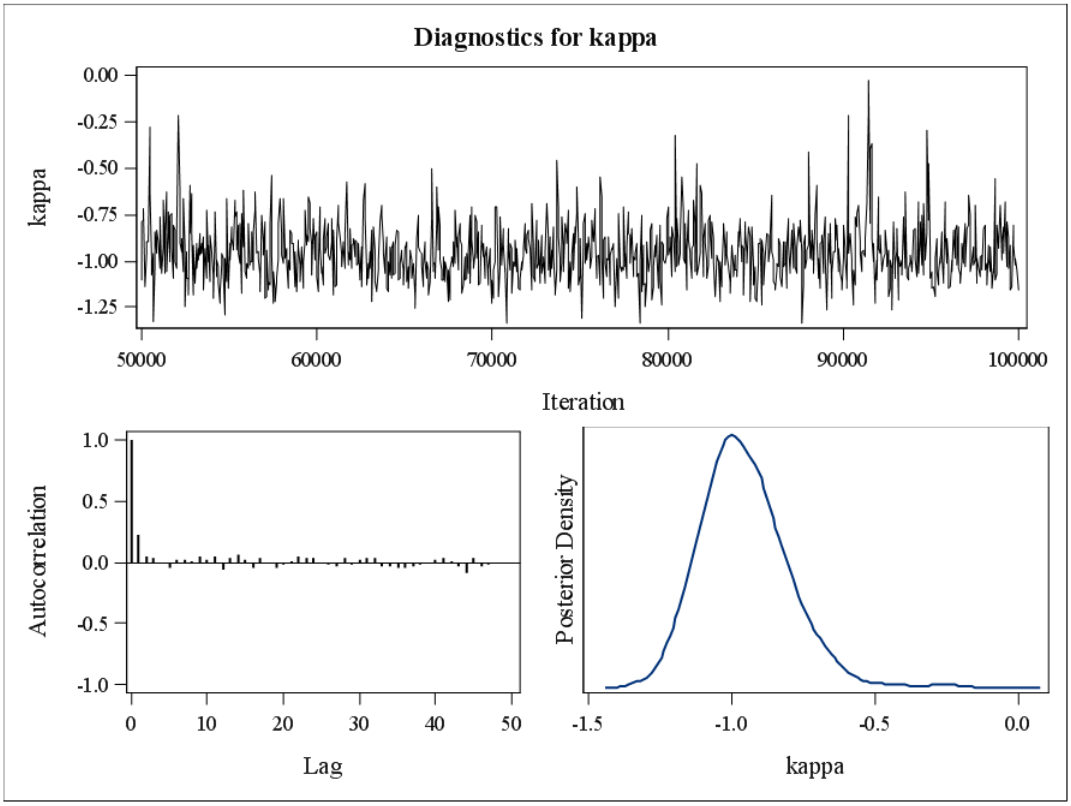


Figure D.19: Convergence diagnostics for the random effects covariance parameter, $\sigma_{11} = \text{Var}(a_{1i})$

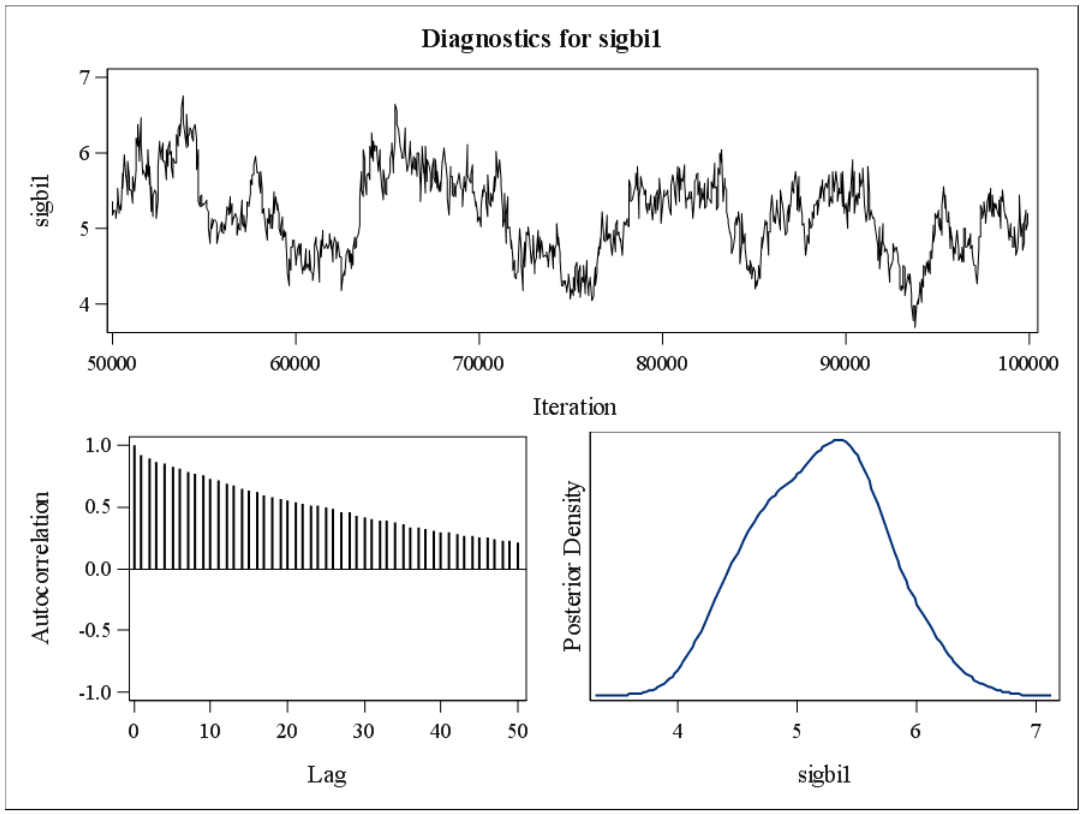


Figure D.20: Convergence diagnostics for the random effects covariance parameter, $\sigma_{22} = \text{Var}(a_{2i})$

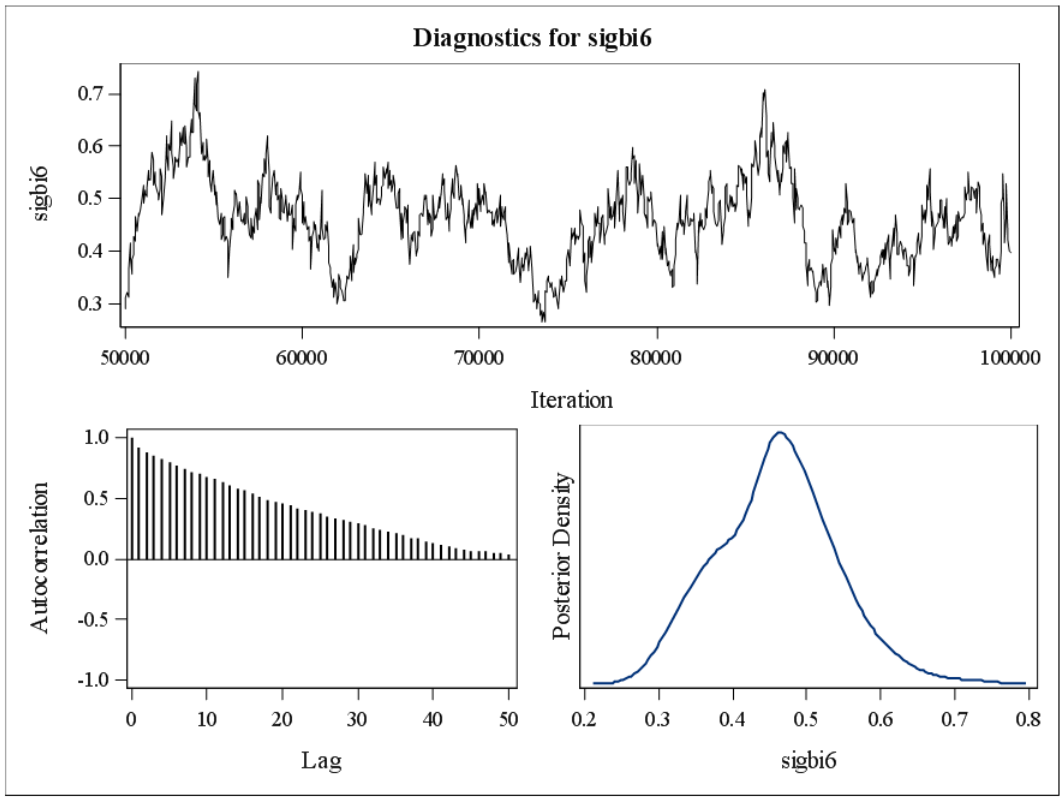


Figure D.21: Convergence diagnostics for the random effects covariance parameter, $\sigma_{33} = \text{Var}(d_{1i})$

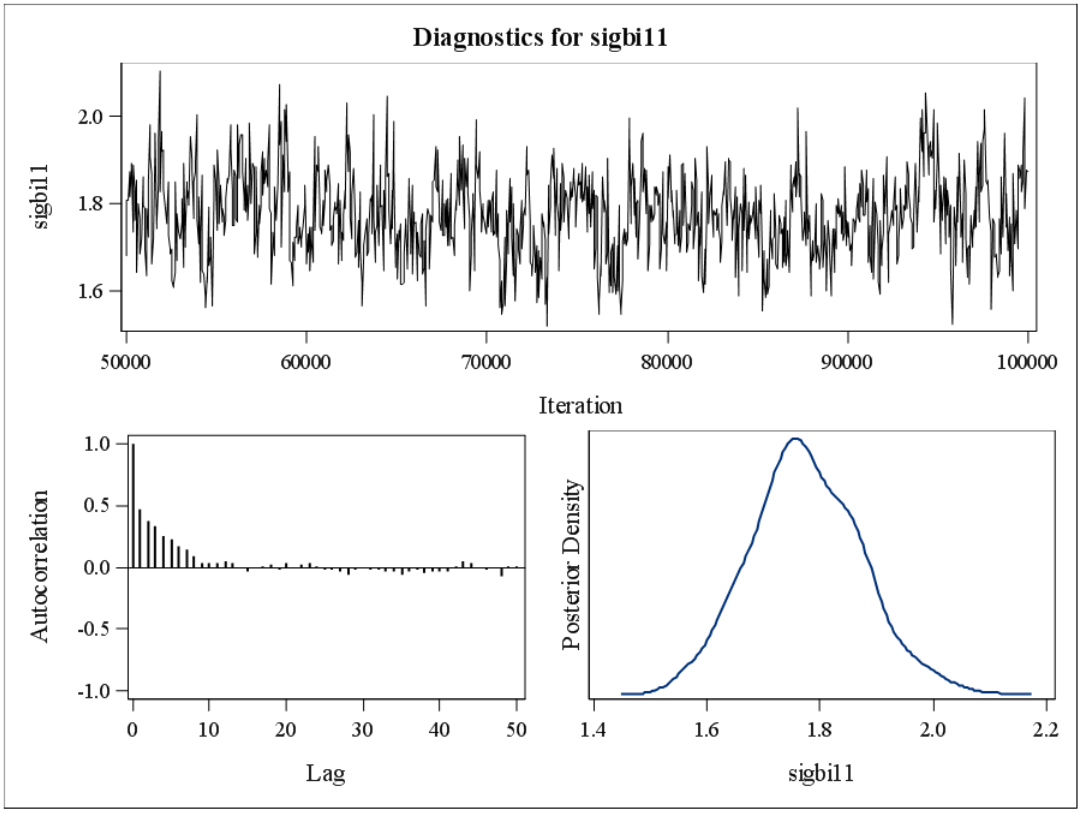


Figure D.22: Convergence diagnostics for the random effects covariance parameter, $\sigma_{44} = \text{Var}(d_{2i})$

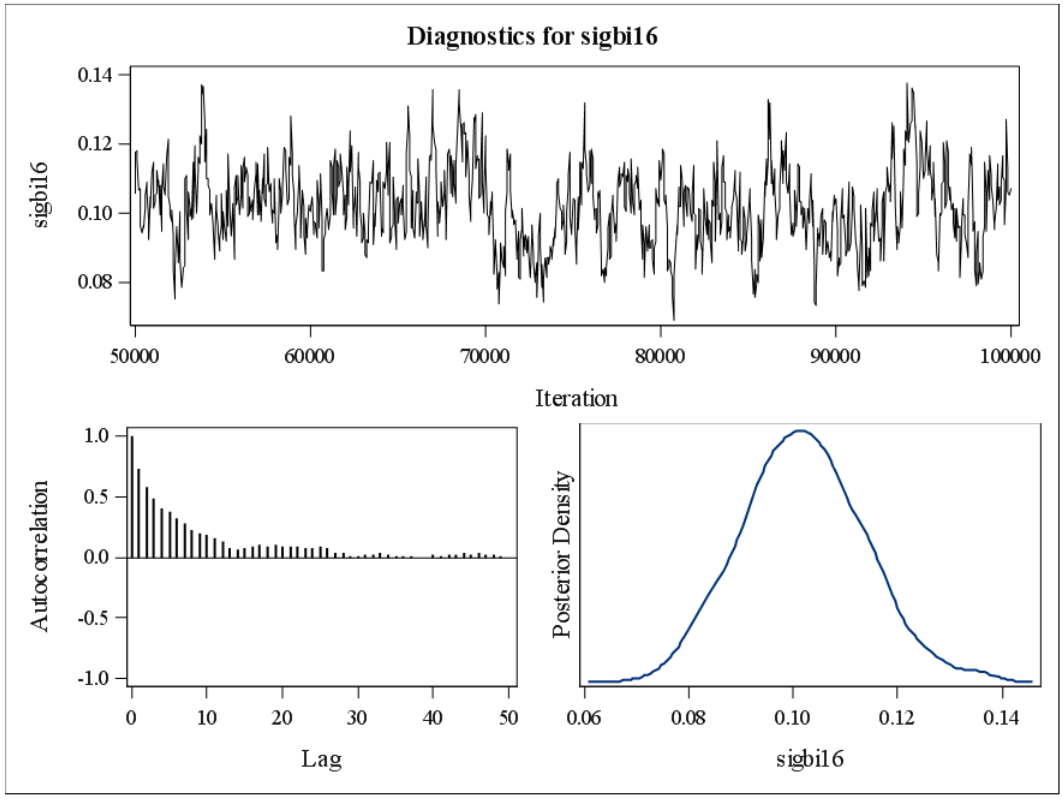


Figure D.23: Convergence diagnostics for the random effects covariance parameter, $\sigma_{12} = \text{Cov}(a_{1i}, a_{2i})$

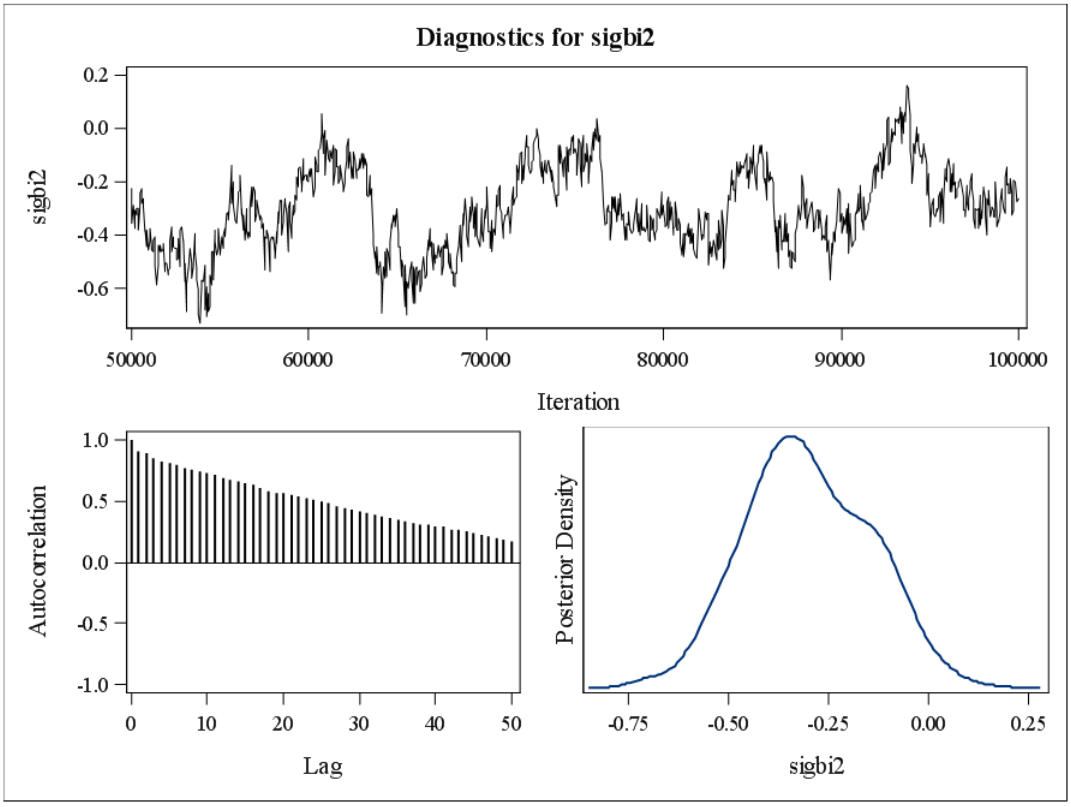


Figure D.24: Convergence diagnostics for the random effects covariance parameter, $\sigma_{13} = \text{Cov}(a_{1i}, d_{1i})$

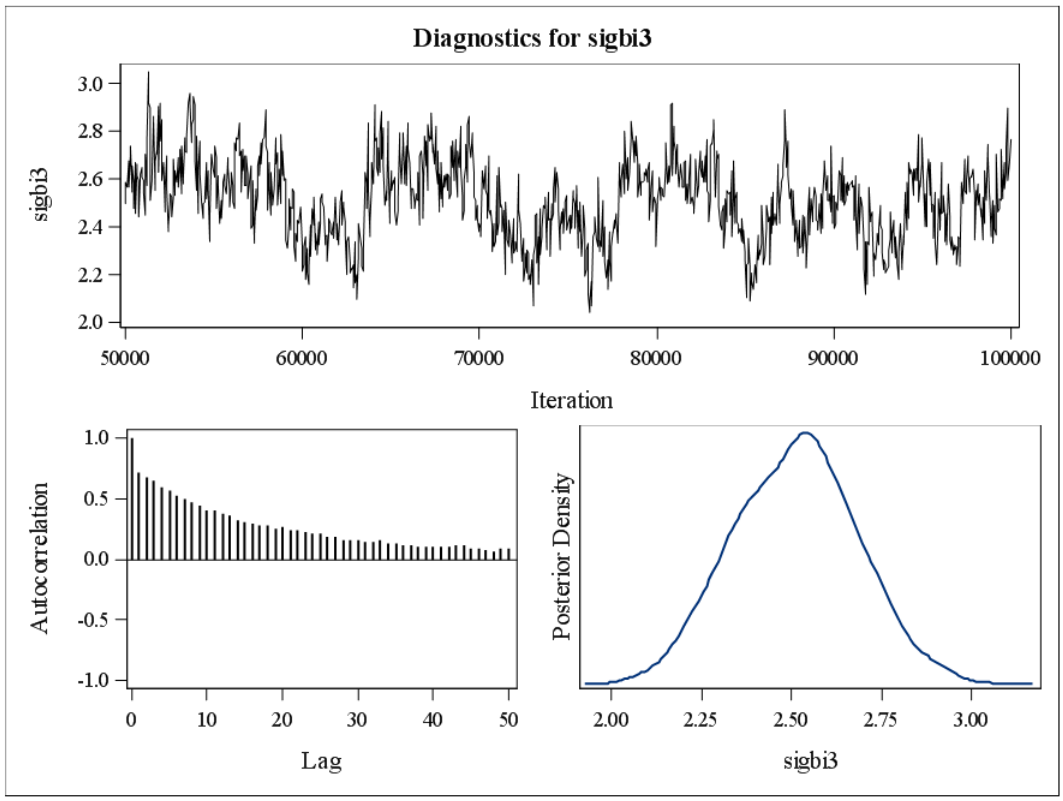


Figure D.25: Convergence diagnostics for the random effects covariance parameter, $\sigma_{14} = \text{Cov}(a_{1i}, d_{2i})$

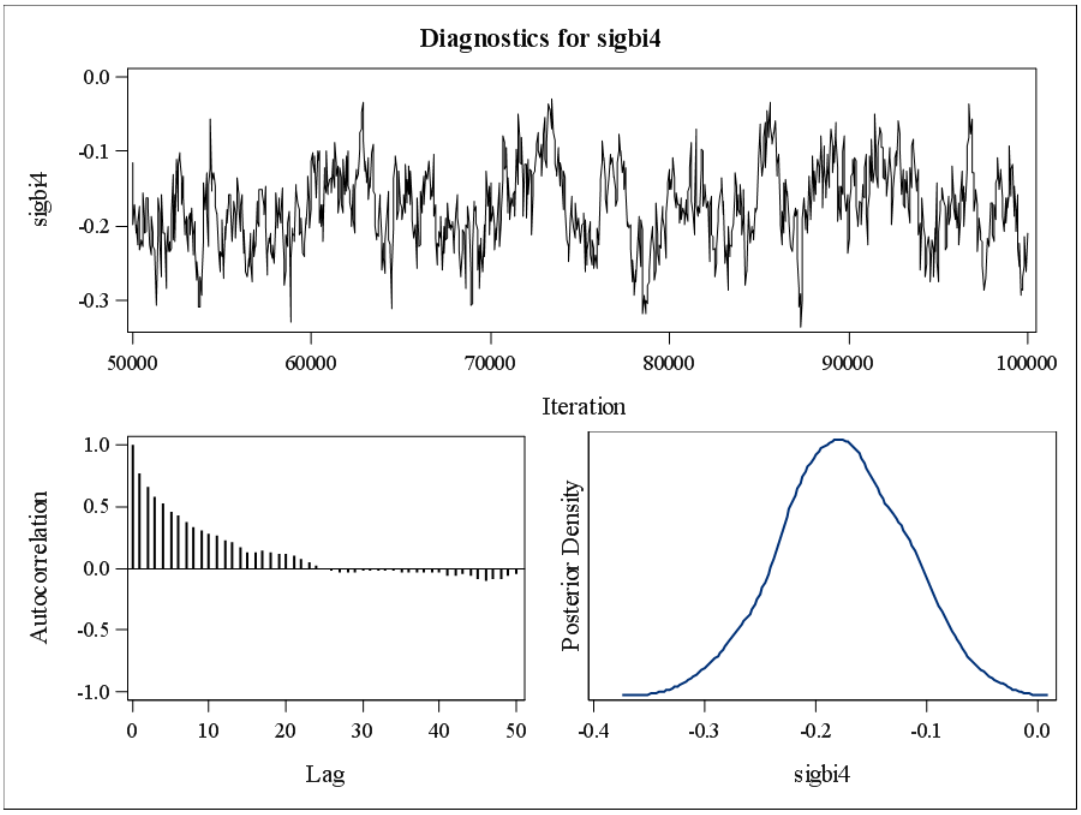


Figure D.26: Convergence diagnostics for the random effects covariance parameter, $\sigma_{23} = \text{Cov}(a_{2i}, d_{1i})$

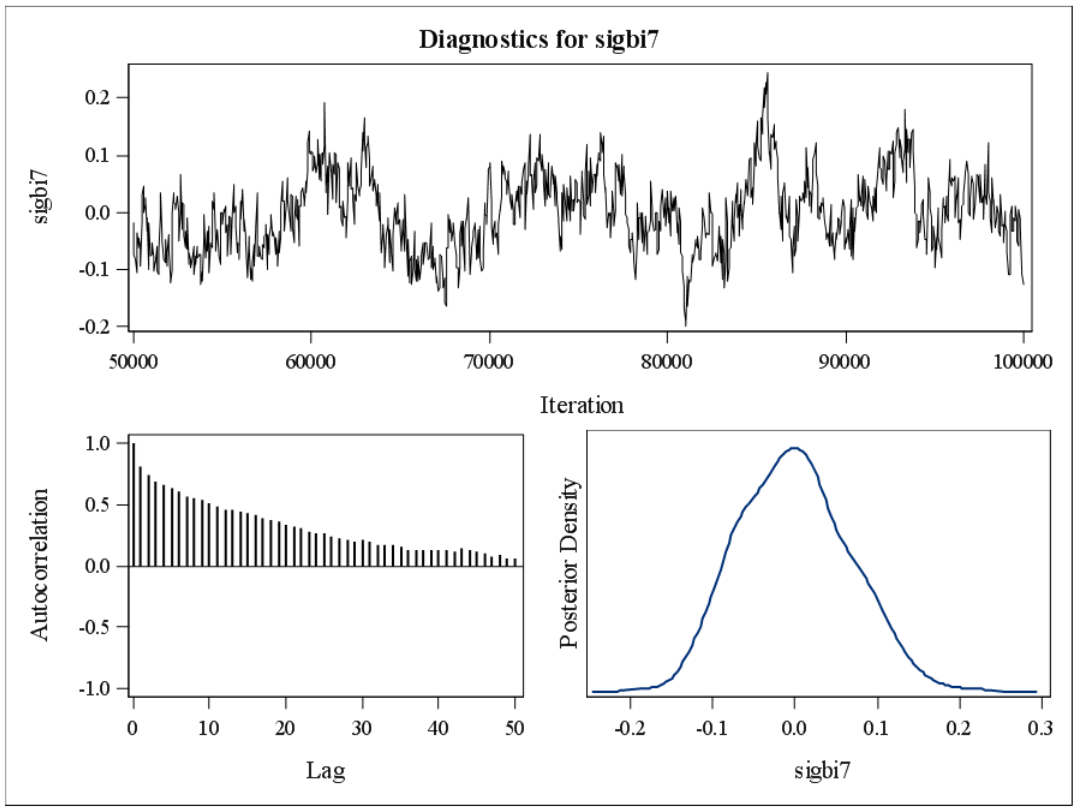


Figure D.27: Convergence diagnostics for the random effects covariance parameter, $\sigma_{24} = \text{Cov}(a_{2i}, d_{2i})$

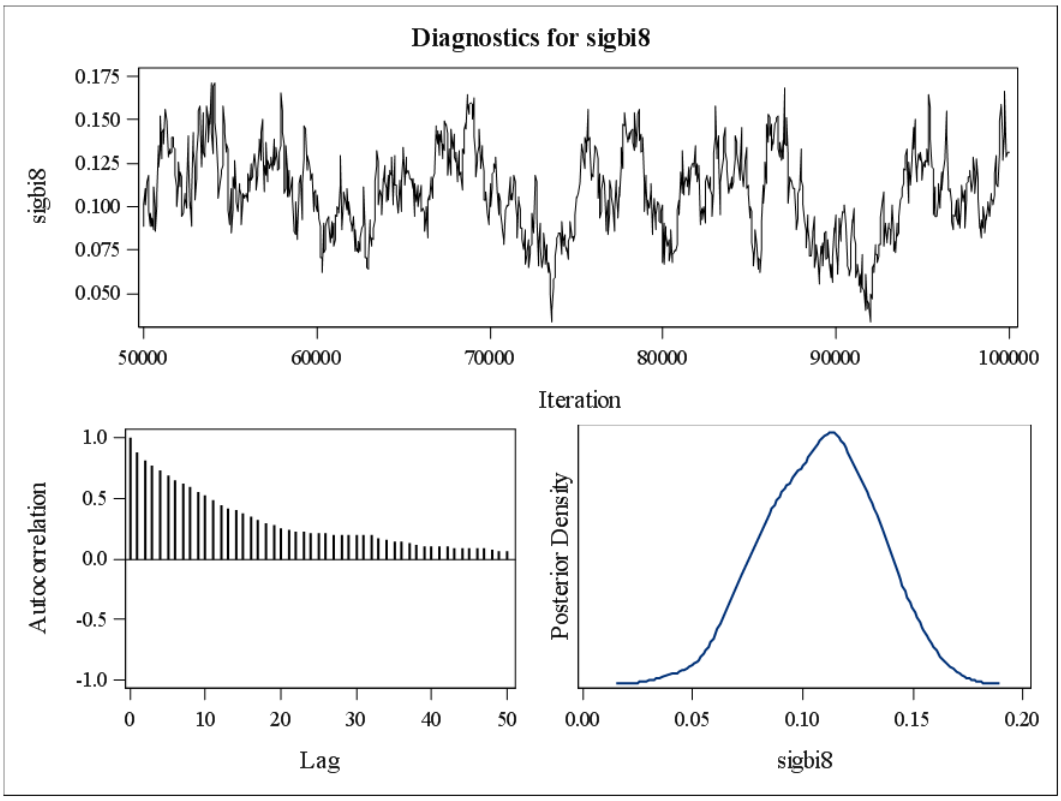
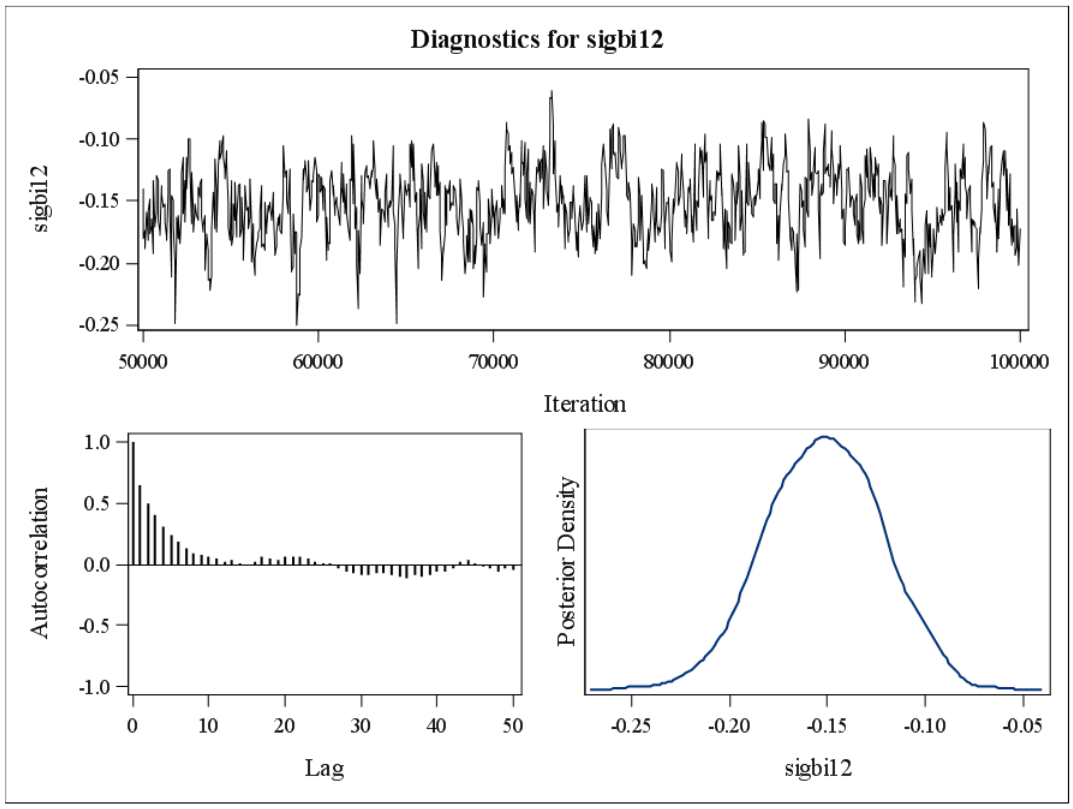


Figure D.28: Convergence diagnostics for the random effects covariance parameter, $\sigma_{34} = \text{Cov}(d_{1i}, d_{2i})$



APPENDIX E: SIMULATION DETAILS FROM CHAPTER 4

Simulation 1a: Data generated as LSN with lower skewness on the log scale and approximately 20% zeros

The simulated data were generated as:

$$\begin{aligned} \text{logit}(\pi_i) &= 3 - 4x_{1i} + 3.5x_{2i} + 2.5x_{3i} \quad \text{and} \\ \text{E}(Y_i) = \nu_i &= \exp(6 + 0.2x_{1i} - 0.01x_{2i} + 0.05x_{3i}) \end{aligned} \tag{E.1}$$

where $x_{1i} \sim$ Bernoulli(0.5), $x_{2i} \sim N(0,1)$ and $x_{3i} \sim$ Pois(1). We generated 1,000 samples of size 200, 1,000, and 10,000 assuming the positive values followed a LSN distribution with $\omega = 1.2$ and $\kappa = 0.5$. Excess zeros were introduced in the Y_i's with probability π_i.

The mean model specification in all four models was $\text{E}(Y_i) = \nu_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})$, and robust sandwich standard errors were used for the GLMs. We examined the bias and coverage probabilities of 95% Wald-type confidence intervals for the parameters included in the overall mean model as well as the prediction of total costs. We assumed that x_{1i} was a binary indicator of treatment arm and the covariate of main interest, and we generated data following the same specification as above in equation (E.1) but with $\beta_1 = 0$ to assess type 1 error at the nominal 0.05 significance level under each model. Results, including mean and median bias, coverage probabilities, and type 1 error rates are included below in Tables E.1, E.2, and E.3 for sample sizes 200, 1,000, and 10,000, respectively.

Table E.1: Model performance on independent outcomes of sample size 200 generated from the model in equation (E.1) with $\kappa = 0.5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0123	-0.0154	-0.3	0.944	
	β_1	0.2	0.0005	0.0009	0.5	0.936	
	β_2	-0.01	-0.0037	-0.0016	15.7	0.945	0.068
	β_3	0.05	-0.0056	-0.0032	-6.4	0.941	
	Total Cost		3.18	-9.47		0.924	
One-part GLM with constant variance	β_0	6	-1.9408	-0.0836	-1.4	0.881	
	β_1	0.2	0.0508	-0.0485	-24.2	0.856	
	β_2	-0.01	-1.0906	0.0611	-611	0.877	0.157
	β_3	0.05	-0.1005	0.0111	22.2	0.884	
	Total Cost		-28.30	-50.19		0.855	
One-part GLM with variance proportional to the mean	β_0	6	-0.0598	-0.0699	-1.2	0.866	
	β_1	0.2	-0.0587	-0.0823	-41.1	0.883	
	β_2	-0.01	0.0609	0.1029	-1029	0.834	0.115
	β_3	0.05	0.0081	0.0286	57.2	0.899	
	Total Cost		-5.28	-35.52		0.857	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0712	-0.0785	-1.3	0.850	
	β_1	0.2	-0.1247	-0.1281	-64.0	0.842	
	β_2	-0.01	0.1505	0.1580	-1580	0.739	0.158
	β_3	0.05	0.0563	0.0586	117	0.810	
	Total Cost		-3.25	-30.44		0.815	

Table E.2: Model performance on independent outcomes of sample size 1,000 generated from the model in equation (E.1) with $\kappa = 0.5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0035	-0.0009	-0.02	0.953	
	β_1	0.2	0.0008	0.0050	2.5	0.932	
	β_2	-0.01	-0.0012	0.0001	-1.1	0.948	0.066
	β_3	0.05	0.0019	0.0027	5.3	0.960	
	Total Cost		1.71	-0.81		0.949	
One-part GLM with constant variance	β_0	6	-3.2312	-0.0426	-0.7	0.883	
	β_1	0.2	0.5217	-0.0128	-6.4	0.904	
	β_2	-0.01	-1.6124	0.0601	-601	0.821	0.100
	β_3	0.05	-0.0954	0.0157	31.4	0.906	
	Total Cost		-13.86	-19.86		0.873	
One-part GLM with variance proportional to the mean	β_0	6	-0.0289	-0.0316	-0.5	0.881	
	β_1	0.2	-0.0133	-0.0282	-14.1	0.903	
	β_2	-0.01	0.0276	0.0665	-665	0.817	0.098
	β_3	0.05	0.0055	0.0217	43.4	0.888	
	Total Cost		-1.86	-14.11		0.875	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0291	-0.0296	-0.5	0.882	
	β_1	0.2	-0.0417	-0.0503	-25.1	0.880	
	β_2	-0.01	0.0650	0.0768	-768	0.756	0.120
	β_3	0.05	0.0240	0.0286	57.1	0.848	
	Total Cost		-2.61	-11.26		0.854	

Table E.3: Model performance on independent outcomes of sample size 10,000 generated from the model in equation (E.1) with $\kappa = 0.5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0007	0.0003	0.005	0.959	
	β_1	0.2	-0.0006	-0.0004	-0.2	0.960	
	β_2	-0.01	0.0003	0.0004	-4.1	0.949	0.042
	β_3	0.05	0	0.00004	0.1	0.946	
	Total Cost		-0.29	-0.26		0.956	
One-part GLM with constant variance	β_0	6	-3.2281	-0.0428	-0.7	0.882	
	β_1	0.2	0.5212	-0.0127	-6.4	0.904	
	β_2	-0.01	-1.6106	0.0601	-601	0.821	0.073
	β_3	0.05	-0.0952	0.0159	31.8	0.906	
	Total Cost		-5.68	-3.86		0.898	
One-part GLM with variance proportional to the mean	β_0	6	-0.0027	-0.0061	-0.1	0.921	
	β_1	0.2	-0.0010	-0.0116	-5.8	0.929	
	β_2	-0.01	0.0016	0.0261	-261	0.830	0.071
	β_3	0.05	-0.0007	0.0068	13.6	0.909	
	Total Cost		0.79	-2.48		0.903	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0024	-0.0052	-0.1	0.926	
	β_1	0.2	-0.0084	-0.0137	-6.9	0.916	
	β_2	-0.01	0.0132	0.0261	-261	0.820	0.084
	β_3	0.05	0.0040	0.0078	15.5	0.892	
	Total Cost		0.43	-1.79		0.896	

Simulation 1b: Data generated as LSN with higher skewness on the log scale and approximately 20% zeros

The simulated data were generated as in equation (E.1) with the log-scale skewness parameter of the LSN distribution set to $\kappa = 5$. Fitting the same models as above, results from these simulated datasets are included below in Tables E.4, E.5, and E.6 for sample sizes 200, 1,000, and 10,000, respectively.

Table E.4: Model performance on independent outcomes of sample size 200 generated from the model in equation (E.1) with $\kappa = 5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0158	-0.0212	-0.4	0.925	
	β_1	0.2	0.0067	0.0055	2.7	0.896	
	β_2	-0.01	-0.0011	0.0004	-4.5	0.908	0.107
	β_3	0.05	0.0003	-0.0013	-2.6	0.891	
	Total Cost		-0.40	-6.98		0.914	
One-part GLM with constant variance	β_0	6	-4.0899	-0.0846	-1.4	0.857	
	β_1	0.2	1.1798	-0.0402	-20.1	0.882	
	β_2	-0.01	-2.2080	0.0823	-823	0.846	0.127
	β_3	0.05	-0.1945	0.0199	39.9	0.895	
	Total Cost		-25.30	-41.06		0.849	
One-part GLM with variance proportional to the mean	β_0	6	-0.0533	-0.0694	-1.2	0.871	
	β_1	0.2	-0.0529	-0.0704	-35.2	0.890	
	β_2	-0.01	0.0601	0.1046	-1046	0.807	0.110
	β_3	0.05	0.0128	0.0309	61.8	0.887	
	Total Cost		-6.21	-29.13		0.853	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0613	-0.0727	-1.2	0.859	
	β_1	0.2	-0.1125	-0.1061	-53.1	0.848	
	β_2	-0.01	0.1376	0.1409	-1409	0.730	0.152
	β_3	0.05	0.0538	0.0497	99.3	0.797	
	Total Cost		-4.77	-24.25		0.816	

Table E.5: Model performance on independent outcomes of sample size 1,000 generated from the model in equation (E.1) with $\kappa = 5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0051	-0.0061	-0.1	0.960	
	β_1	0.2	0.0023	0.0011	0.5	0.945	
	β_2	-0.01	0.0018	0.0007	-6.9	0.934	0.055
	β_3	0.05	0.0007	0.0007	1.4	0.961	
	Total Cost		-0.58	-1.63		0.950	
One-part GLM with constant variance	β_0	6	-2.4745	-0.0415	-0.7	0.892	
	β_1	0.2	0.5008	-0.0180	-9.0	0.906	
	β_2	-0.01	-1.1942	0.0563	-563	0.805	0.098
	β_3	0.05	-0.1091	0.0159	31.9	0.898	
	Total Cost		-12.21	-15.99		0.874	
One-part GLM with variance proportional to the mean	β_0	6	-0.0232	-0.0294	-0.5	0.896	
	β_1	0.2	-0.0123	-0.0356	-17.8	0.909	
	β_2	-0.01	0.0231	0.0629	-629	0.794	0.090
	β_3	0.05	0.0057	0.0195	38.9	0.881	
	Total Cost		-1.76	-11.23		0.875	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0230	-0.0277	-0.5	0.895	
	β_1	0.2	-0.0369	-0.0510	-25.5	0.884	
	β_2	-0.01	0.0560	0.0722	-722	0.760	0.116
	β_3	0.05	0.0210	0.0267	53.3	0.838	
	Total Cost		-2.35	-8.76		0.857	

Table E.6: Model performance on independent outcomes of sample size 10,000 generated from the model in equation (E.1) with $\kappa = 5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0008	-0.0005	-0.008	0.956	
	β_1	0.2	-0.0013	-0.0018	-0.9	0.955	
	β_2	-0.01	0.0002	-0.0002	1.7	0.952	0.045
	β_3	0.05	0.0003	0.0003	0.6	0.945	
	Total Cost		-0.47	-0.59		0.953	
One-part GLM with constant variance	β_0	6	-0.1478	-0.0079	-0.1	0.916	
	β_1	0.2	0.0214	-0.0075	-3.7	0.911	
	β_2	-0.01	-0.0794	0.0231	-231	0.818	0.092
	β_3	0.05	-0.0362	0.0052	10.5	0.908	
	Total Cost		-5.92	-3.15		0.896	
One-part GLM with variance proportional to the mean	β_0	6	-0.0027	-0.0061	-0.1	0.925	
	β_1	0.2	-0.0015	-0.0096	-4.8	0.916	
	β_2	-0.01	0.0019	0.0231	-231	0.823	0.083
	β_3	0.05	-0.0001	0.0060	12.0	0.907	
	Total Cost		0.32	-2.04		0.902	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0023	-0.0052	-0.1	0.925	
	β_1	0.2	-0.0077	-0.0116	-5.8	0.906	
	β_2	-0.01	0.0115	0.0235	-235	0.814	0.094
	β_3	0.05	0.0036	0.0065	13.0	0.893	
	Total Cost		0.04	-1.51		0.896	

Simulation 1c: Data generated as LSN with lower skewness on the log scale and approximately 40% zeros

The simulated data were generated as:

$$\begin{aligned} \text{logit}(\pi_i) &= 3 - 7x_{1i} + 5x_{2i} + 2x_{3i} \quad \text{and} \\ \text{E}(Y_i) &= \nu_i = \exp(6 + 0.2x_{1i} - 0.01x_{2i} + 0.05x_{3i}) \end{aligned} \tag{E.2}$$

where, as above, $x_{1i} \sim$ Bernoulli(0.5), $x_{2i} \sim N(0, 1)$ and $x_{3i} \sim$ Pois(1). We again generated 1,000 samples of size 200, 1,000, and 10,000 assuming the positive values followed a LSN distribution with $\omega = 1.2$ and $\kappa = 0.5$. Excess zeros were introduced in the Y_i's with probability π_i. Note that the overall mean model remained the same. Results from fitting the four models to these datasets are included below in Tables E.7, E.8, and E.9 for sample sizes 200, 1,000, and 10,000, respectively.

Table E.7: Model performance on independent outcomes of sample size 200 generated from the model in equation (E.2) with $\kappa = 0.5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0182	-0.0230	-0.4	0.936	
	β_1	0.2	-0.0031	0.0084	4.2	0.939	
	β_2	-0.01	-0.0013	-0.0011	10.6	0.944	0.062
	β_3	0.05	-0.0031	-0.0017	-3.3	0.934	
	Total Cost		10.38	-9.20		0.907	
One-part GLM with constant variance	β_0	6	-2.3632	-0.1941	-3.2	0.806	
	β_1	0.2	-0.2215	-0.2905	-145	0.684	
	β_2	-0.01	-0.7365	0.2532	-2532	0.682	0.328
	β_3	0.05	-0.0106	0.0502	100	0.805	
	Total Cost		-17.83	-112.26		0.704	
One-part GLM with variance proportional to the mean	β_0	6	-0.1560	-0.1813	-3.0	0.788	
	β_1	0.2	-0.3316	-0.4307	-215	0.673	
	β_2	-0.01	0.2861	0.3485	-3485	0.555	0.325
	β_3	0.05	0.0321	0.0889	178	0.790	
	Total Cost		0.12	-95.16		0.696	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.1314	-0.1462	-2.4	0.833	
	β_1	0.2	-0.9569	-0.9077	-454	0.482	
	β_2	-0.01	0.8759	0.7910	-7910	0.334	0.518
	β_3	0.05	0.2675	0.2445	489	0.566	
	Total Cost		330.17	-82.25		0.583	

Table E.8: Model performance on independent outcomes of sample size 1,000 generated from the model in equation (E.2) with $\kappa = 0.5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0053	-0.0022	-0.04	0.951	
	β_1	0.2	0.0043	0.0105	5.2	0.956	
	β_2	-0.01	-0.0046	-0.0022	22.5	0.956	0.044
	β_3	0.05	0.0020	0.0032	6.3	0.952	
	Total Cost		3.28	0.36		0.951	
One-part GLM with constant variance	β_0	6	-0.5069	-0.1488	-2.5	0.764	
	β_1	0.2	-0.1279	-0.1600	-80.0	0.746	
	β_2	-0.01	-0.0075	0.2118	-2118	0.553	0.265
	β_3	0.05	0.0172	0.0625	125	0.776	
	Total Cost		-25.63	-64.72		0.718	
One-part GLM with variance proportional to the mean	β_0	6	-0.1242	-0.1407	-2.3	0.740	
	β_1	0.2	-0.1796	-0.2611	-131	0.695	
	β_2	-0.01	0.2099	0.2786	-2786	0.472	0.305
	β_3	0.05	0.0549	0.0864	173	0.720	
	Total Cost		-19.75	-58.91		0.683	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.1027	-0.1126	-1.9	0.817	
	β_1	0.2	-0.4380	-0.4651	-233	0.552	
	β_2	-0.01	0.4502	0.4750	-4750	0.351	0.448
	β_3	0.05	0.1508	0.1599	320	0.579	
	Total Cost		0.71	-48.93		0.611	

Table E.9: Model performance on independent outcomes of sample size 10,000 generated from the model in equation (E.2) with $\kappa = 0.5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	0.0008	0.0009	0.01	0.958	
	β_1	0.2	-0.0008	-0.0013	-0.7	0.954	
	β_2	-0.01	0.00004	-0.0003	2.9	0.948	0.047
	β_3	0.05	0.0001	0.0001	0.3	0.947	
	Total Cost		0.62	0.42		0.952	
One-part GLM with constant variance	β_0	6	-0.1088	-0.0680	-1.1	0.737	
	β_1	0.2	-0.0202	-0.0938	-46.9	0.751	
	β_2	-0.01	0.0246	0.1353	-1353	0.557	0.256
	β_3	0.05	-0.0532	0.0365	73.0	0.751	
	Total Cost		-11.22	-26.07		0.723	
One-part GLM with variance proportional to the mean	β_0	6	-0.0418	-0.0600	-1.0	0.760	
	β_1	0.2	-0.0658	-0.1192	-59.6	0.726	
	β_2	-0.01	0.0933	0.1574	-1574	0.544	0.274
	β_3	0.05	0.0184	0.0454	90.8	0.728	
	Total Cost		-5.70	-20.02		0.714	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0317	-0.0492	-0.8	0.838	
	β_1	0.2	-0.1407	-0.1686	-84.3	0.664	
	β_2	-0.01	0.1708	0.2023	-2203	0.485	0.335
	β_3	0.05	0.0485	0.0601	120	0.683	
	Total Cost		-7.74	-14.49		0.687	

Simulation 1d: Data generated as LSN with higher skewness on the log scale and approximately 40% zeros

The simulated data were generated as in equation (E.2) with the log-scale skewness parameter of the LSN distribution set to $\kappa = 5$. Fitting the same models as above, results from these simulated datasets are included below in Tables E.10, E.11, and E.12 for sample sizes 200, 1,000, and 10,000, respectively.

Table E.10: Model performance on independent outcomes of sample size 200 generated from the model in equation (E.2) with $\kappa = 5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0260	-0.0212	-0.4	0.905	
	β_1	0.2	0.0199	0.0210	10.5	0.861	
	β_2	-0.01	0.0028	-0.0017	17.1	0.883	0.145
	β_3	0.05	0.0021	-0.0010	-2.0	0.864	
	Total Cost		3.44	-6.99		0.878	
One-part GLM with constant variance	β_0	6	-1.7351	-0.1764	-2.9	0.793	
	β_1	0.2	-0.1327	-0.2510	-125	0.689	
	β_2	-0.01	-0.6421	0.2619	-2619	0.632	0.325
	β_3	0.05	-0.0637	0.0509	102	0.810	
	Total Cost		-14.85	-100.15		0.700	
One-part GLM with variance proportional to the mean	β_0	6	-0.1463	-0.1613	-2.7	0.780	
	β_1	0.2	-0.3000	-0.4135	-207	0.661	
	β_2	-0.01	0.2711	0.3549	-3549	0.506	0.338
	β_3	0.05	0.0373	0.0864	173	0.779	
	Total Cost		0.64	-85.52		0.688	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.1177	-0.1329	-2.2	0.831	
	β_1	0.2	-0.8678	-0.8546	-427	0.480	
	β_2	-0.01	0.8051	0.7658	-7658	0.317	0.520
	β_3	0.05	0.2485	0.2296	459	0.577	
	Total Cost		218.24	-72.19		0.581	

Table E.11: Model performance on independent outcomes of sample size 1,000 generated from the model in equation (E.2) with $\kappa = 5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0067	-0.0081	-0.1	0.959	
	β_1	0.2	0.0077	0.0065	3.3	0.936	
	β_2	-0.01	-0.0002	-0.0008	8.2	0.934	0.064
	β_3	0.05	0.0009	0.0011	2.2	0.957	
	Total Cost		0.81	-0.83		0.948	
One-part GLM with constant variance	β_0	6	-0.8801	-0.1310	-2.2	0.723	
	β_1	0.2	-0.1091	-0.1568	-78.4	0.755	
	β_2	-0.01	-0.2470	0.2050	-2050	0.519	0.261
	β_3	0.05	0.0591	0.0621	124	0.755	
	Total Cost		-24.94	-58.26		0.711	
One-part GLM with variance proportional to the mean	β_0	6	-0.1087	-0.1214	-2.0	0.728	
	β_1	0.2	-0.1588	-0.2284	-114	0.700	
	β_2	-0.01	0.1910	0.2666	-2666	0.461	0.301
	β_3	0.05	0.0496	0.0808	162	0.706	
	Total Cost		-17.68	-51.41		0.685	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0904	-0.1010	-1.7	0.804	
	β_1	0.2	-0.3835	-0.4029	-201	0.584	
	β_2	-0.01	0.4040	0.4286	-4286	0.364	0.416
	β_3	0.05	0.1351	0.1384	277	0.582	
	Total Cost		-1.89	-41.19		0.623	

Table E.12: Model performance on independent outcomes of sample size 10,000 generated from the model in equation (E.2) with $\kappa = 5$ and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0010	-0.0012	-0.02	0.959	0.052
	β_1	0.2	-0.0007	-0.0020	-1.0	0.948	
	β_2	-0.01	0.00005	0.00005	-0.5	0.951	
	β_3	0.05	0.0004	0.0004	0.9	0.931	
	Total Cost		-0.30	-0.43		0.952	
One-part GLM with constant variance	β_0	6	-0.1016	-0.0633	-1.1	0.734	0.265
	β_1	0.2	-0.0177	-0.0823	-41.2	0.742	
	β_2	-0.01	0.0169	0.1280	-1280	0.561	
	β_3	0.05	-0.0489	0.0342	68.5	0.746	
	Total Cost		-11.16	-23.47		0.720	
One-part GLM with variance proportional to the mean	β_0	6	-0.0398	-0.0541	-0.9	0.755	0.284
	β_1	0.2	-0.0582	-0.1078	-53.9	0.716	
	β_2	-0.01	0.0839	0.1474	-1474	0.551	
	β_3	0.05	0.0183	0.0413	82.6	0.738	
	Total Cost		-5.28	-17.59		0.713	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0287	-0.0410	-0.7	0.825	0.336
	β_1	0.2	-0.1250	-0.1475	-73.7	0.664	
	β_2	-0.01	0.1532	0.1802	-1802	0.518	
	β_3	0.05	0.0444	0.0542	108	0.684	
	Total Cost		-6.43	-11.96		0.691	

Simulation 2a: Data generated as generalized gamma distributed and approximately 20% zeros

As above, the simulated data were generated under the model:

$$\begin{aligned} \text{logit}(\pi_i) &= 3 - 4x_{1i} + 3.5x_{2i} + 2.5x_{3i} \quad \text{and} \\ \text{E}(Y_i) &= \nu_i = \exp(6 + 0.2x_{1i} - 0.01x_{2i} + 0.05x_{3i}) \end{aligned} \tag{E.3}$$

where $x_{1i} \sim$ Bernoulli(0.5), $x_{2i} \sim N(0,1)$ and $x_{3i} \sim$ Pois(1). We generated 1,000 samples of size 200, 1,000, and 10,000 assuming the positive values followed a generalized gamma distribution with $\sigma = 1.2$ and $\kappa = 0.63$. This is following the parameterization used in Liu et al. (2010) with

$$f(y_i; \kappa, \mu_i, \sigma) = \frac{\eta^\eta}{\sigma y_i \Gamma(\eta)\sqrt{\eta}} \exp\left[u_i\sqrt{\eta} - \eta\exp(|\kappa|u_i)\right],$$

where $\eta = |\kappa|^{-2}$, $u_i = \text{sign}(\kappa)\left(\log(y_i) - \mu_i\right)/\sigma$, μ_i is the location parameter, $\sigma > 0$ is the scale parameter, and κ is the shape parameter. We then have

$$\begin{aligned} \text{E}(Y_i) = \nu_i &= \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}\right) \\ &= \pi_i \exp\left\{\mu_i + \frac{\sigma\log(\kappa^2)}{\kappa} + \log\left[\Gamma\left(1/\kappa^2 + \sigma/\kappa\right)\right] - \log\left[\Gamma\left(1/\kappa^2\right)\right]\right\}. \end{aligned}$$

Solving for μ_i in terms of ν_i, we obtain

$$\begin{aligned} \mu_i &= \log(\nu_i) - \log(\pi_i) - \frac{\sigma\log(\kappa^2)}{\kappa} - \log\left[\Gamma\left(1/\kappa^2 + \sigma/\kappa\right)\right] + \log\left[\Gamma\left(1/\kappa^2\right)\right] \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} - \log(\pi_i) - \frac{\sigma\log(\kappa^2)}{\kappa} - \log\left[\Gamma\left(1/\kappa^2 + \sigma/\kappa\right)\right] + \log\left[\Gamma\left(1/\kappa^2\right)\right], \end{aligned}$$

and use this form to generate the positive values of y_i with $\sigma = 1.2$ and $\kappa = 0.63$. Excess zeros were introduced in the Y_i's with probability π_i. Fitting the same models as above, results from these simulated datasets are included below in Tables E.13, E.14, and E.15 for sample sizes 200, 1,000, and 10,000, respectively.

Table E.13: Model performance on independent outcomes of sample size 200 generated from the model in equation (E.3) under the generalized gamma distribution and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	0.0114	0.0170	0.3	0.946	
	β_1	0.2	-0.0053	-0.0080	-4.0	0.944	
	β_2	-0.01	-0.0038	-0.0032	-32.2	0.940	0.054
	β_3	0.05	-0.0048	-0.0035	-6.9	0.928	
	Total Cost		14.16	3.18		0.929	
One-part GLM with constant variance	β_0	6	-1.4756	-0.0775	-1.3	0.917	
	β_1	0.2	-0.0342	-0.0312	-15.6	0.870	
	β_2	-0.01	-1.0509	0.0733	-733	0.864	0.133
	β_3	0.05	-0.1031	0.0137	27.4	0.877	
	Total Cost		-14.89	-35.57		0.873	
One-part GLM with variance proportional to the mean	β_0	6	-0.0487	-0.0526	-0.9	0.905	
	β_1	0.2	-0.0416	-0.0737	-36.9	0.887	
	β_2	-0.01	0.0471	0.0906	-906	0.824	0.113
	β_3	0.05	0.0068	0.0271	54.2	0.880	
	Total Cost		3.24	-24.04		0.876	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0574	-0.0610	-1.0	0.890	
	β_1	0.2	-0.1080	-0.1117	-55.9	0.851	
	β_2	-0.01	0.1362	0.1376	-1376	0.735	0.149
	β_3	0.05	0.0560	0.0527	105	0.797	
	Total Cost		3.48	-19.35		0.834	

Table E.14: Model performance on independent outcomes of sample size 1,000 generated from the model in equation (E.3) under the generalized gamma distribution and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	0.0340	0.0227	0.4	0.863	
	β_1	0.2	0.0047	0.0046	2.3	0.956	
	β_2	-0.01	0.0005	0.0019	-19.1	0.944	0.047
	β_3	0.05	0.0001	0.0012	2.4	0.945	
	Total Cost		21.53	11.44		0.893	
One-part GLM with constant variance	β_0	6	-3.5240	-0.0362	-0.6	0.906	
	β_1	0.2	0.2910	-0.0201	-10.1	0.905	
	β_2	-0.01	-1.7346	0.0551	-551	0.808	0.099
	β_3	0.05	-0.2322	0.0155	31.0	0.901	
	Total Cost		-12.90	-14.14		0.884	
One-part GLM with variance proportional to the mean	β_0	6	-0.0234	-0.0288	-0.5	0.912	
	β_1	0.2	-0.0108	-0.0289	-14.5	0.916	
	β_2	-0.01	0.0254	0.0625	-625	0.797	0.084
	β_3	0.05	0.0049	0.0214	42.9	0.906	
	Total Cost		-0.82	-9.62		0.888	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0233	-0.0312	-0.5	0.915	
	β_1	0.2	-0.0372	-0.0418	-20.9	0.896	
	β_2	-0.01	0.0600	0.0750	-750	0.756	0.104
	β_3	0.05	0.0218	0.0293	58.5	0.863	
	Total Cost		-1.72	-7.36		0.869	

Table E.15: Model performance on independent outcomes of sample size 10,000 generated from the model in equation (E.3) under the generalized gamma distribution and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	0.1197	0.1673	2.8	0.348	0.044
	β_1	0.2	-0.0018	-0.0015	-0.8	0.955	
	β_2	-0.01	0.0008	0.0011	-11.0	0.960	
	β_3	0.05	0.0009	0.0009	1.9	0.952	
	Total Cost		62.28	79.86		0.353	
One-part GLM with constant variance	β_0	6	-0.1629	-0.0093	-0.2	0.903	0.097
	β_1	0.2	0.0545	-0.0109	-5.5	0.907	
	β_2	-0.01	-0.0735	0.0301	-301	0.806	
	β_3	0.05	-0.0385	0.0063	12.6	0.889	
	Total Cost		-4.67	-3.55		0.886	
One-part GLM with variance proportional to the mean	β_0	6	-0.0092	-0.0063	-0.1	0.905	0.101
	β_1	0.2	-0.0044	-0.0130	-6.5	0.899	
	β_2	-0.01	0.0022	0.0301	-301	0.802	
	β_3	0.05	0.0012	0.0071	14.2	0.882	
	Total Cost		1.10	-2.36		0.887	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0047	-0.0049	-0.1	0.916	0.111
	β_1	0.2	-0.0121	-0.0151	-7.5	0.889	
	β_2	-0.01	0.0165	0.0309	-309	0.793	
	β_3	0.05	0.0062	0.0077	15.4	0.871	
	Total Cost		-0.63	-1.67		0.882	

Simulation 2b: Data generated as generalized gamma distributed and approximately 40% zeros

The simulated data were generated as:

$$\begin{aligned} \text{logit}(\pi_i) &= 3 - 7x_{1i} + 5x_{2i} + 2x_{3i} \quad \text{and} \\ \text{E}(Y_i) &= \nu_i = \exp(6 + 0.2x_{1i} - 0.01x_{2i} + 0.05x_{3i}) \end{aligned} \tag{E.4}$$

with all covariates as defined above. Fitting the same models as above, results from these simulated datasets are included below in Tables E.16, E.17, and E.18 for sample sizes 200, 1,000, and 10,000, respectively.

Table E.16: Model performance on independent outcomes of sample size 200 generated from the model in equation (E.4) under the generalized gamma distribution and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	0.0167	0.0184	0.3	0.938	0.057
	β_1	0.2	-0.0070	0.0039	2.0	0.944	
	β_2	-0.01	-0.0024	0.0091	-90.6	0.940	
	β_3	0.05	-0.0039	-0.0035	-7.0	0.924	
	Total Cost		28.02	6.94		0.926	
One-part GLM with constant variance	β_0	6	-4.7434	-0.1700	-2.8	0.828	0.304
	β_1	0.2	-0.4318	-0.2422	-121	0.710	
	β_2	-0.01	-2.1603	0.2321	-2321	0.669	
	β_3	0.05	0.5211	0.0621	124	0.792	
	Total Cost		24.74	-101.05		0.720	
One-part GLM with variance proportional to the mean	β_0	6	-0.1627	-0.1560	-2.6	0.823	0.316
	β_1	0.2	-0.3136	-0.4169	-208	0.686	
	β_2	-0.01	0.2759	0.3562	-3562	0.545	
	β_3	0.05	0.0556	0.0931	186	0.777	
	Total Cost		26.61	-86.45		0.708	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.1292	-0.1181	-2.0	0.848	0.524
	β_1	0.2	-0.9536	-0.8879	-444	0.475	
	β_2	-0.01	0.8648	0.8263	-8263	0.322	
	β_3	0.05	0.2847	0.2509	502	0.560	
	Total Cost		3090.55	-74.79		0.581	

Table E.17: Model performance on independent outcomes of sample size 1,000 generated from the model in equation (E.4) under the generalized gamma distribution and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	-0.0036	-0.0033	-0.1	0.943	
	β_1	0.2	-0.0006	0.0016	0.8	0.948	
	β_2	-0.01	-0.0005	-0.0029	29.0	0.941	0.051
	β_3	0.05	0.0013	0.0009	1.9	0.945	
	Total Cost		3.17	-1.99		0.937	
One-part GLM with constant variance	β_0	6	-2.0154	-0.1289	-2.1	0.756	
	β_1	0.2	0.0340	-0.1841	-92.1	0.715	
	β_2	-0.01	-0.6988	0.2126	-2126	0.512	0.304
	β_3	0.05	-0.0070	0.0564	113	0.750	
	Total Cost		-9.25	-60.39		0.697	
One-part GLM with variance proportional to the mean	β_0	6	-0.0921	-0.1225	-2.0	0.734	
	β_1	0.2	-0.1658	-0.2738	-137	0.667	
	β_2	-0.01	0.2024	0.2803	-2803	0.448	0.333
	β_3	0.05	0.0315	0.0776	155	0.705	
	Total Cost		-1.64	-51.04		0.670	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0777	-0.0986	-1.6	0.825	
	β_1	0.2	-0.4146	-0.4574	-229	0.538	
	β_2	-0.01	0.4309	0.4688	-4688	0.350	0.462
	β_3	0.05	0.1317	0.1456	291	0.596	
	Total Cost		10.09	-38.35		0.610	

Table E.18: Model performance on independent outcomes of sample size 10,000 generated from the model in equation (E.4) under the generalized gamma distribution and 1,000 simulations

Model	Parameter	True Value	Mean Bias	Median Bias	Percent Relative Median Bias	Coverage Probability	Type 1 Error
LSN MTP	β_0	6	0.0862	0.0394	0.7	0.509	
	β_1	0.2	-0.0018	-0.0030	-1.5	0.938	
	β_2	-0.01	-0.0002	0	-0.04	0.951	0.055
	β_3	0.05	0.0009	0.0014	2.8	0.936	
	Total Cost		45.08	27.14		0.539	
One-part GLM with constant variance	β_0	6	-0.7292	-0.0732	-1.2	0.703	
	β_1	0.2	-0.1119	-0.0950	-47.5	0.738	
	β_2	-0.01	-0.2084	0.1430	-1430	0.527	0.264
	β_3	0.05	-0.0569	0.0384	76.8	0.724	
	Total Cost		-18.74	-27.30		0.702	
One-part GLM with variance proportional to the mean	β_0	6	-0.0478	-0.0640	-1.1	0.725	
	β_1	0.2	-0.0760	-0.1243	-62.1	0.719	
	β_2	-0.01	0.1068	0.1693	-1693	0.507	0.280
	β_3	0.05	0.0253	0.0465	92.9	0.693	
	Total Cost		-11.66	-22.23		0.693	
One-part GLM with standard deviation proportional to the mean	β_0	6	-0.0370	-0.0493	-0.8	0.795	
	β_1	0.2	-0.1499	-0.1736	-86.8	0.664	
	β_2	-0.01	0.1827	0.2091	-2091	0.464	0.336
	β_3	0.05	0.0530	0.0615	123	0.651	
	Total Cost		-10.99	-16.72		0.666	

BIBLIOGRAPHY

Aitchison, J. (1955), "On the distribution of a positive random variable having a discrete probability mass at the origin," *Journal of the American Statistical Association*, 50, 901–908.

Albert, P. S. (2005), "Letter to the editor," *Biometrics*, 47, 879–881.

Arterburn, D. E., Maciejewski, M. L., and Tsevat, J. (2005), "Impact of morbid obesity on medical expenditures in adults," *International Journal of Obesity Related Metabolic Disorders*, 29, 334–339.

Azzalini, A. (1985), "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics*, 12, 171–178.

— (2013), *R package* `sn`*: The skew-normal and skew-t distributions (version 0.4-18)*, Università di Padova, Italia.

Basu, A. and Rathouz, P. J. (2005), "Estimating marginal and incremental effects on health outcomes using flexible link and variance function models," *Biostatistics*, 6, 93–109.

Berndt, E. R., Bir, A., Busch, S. H., Frank, R. G., and Normand, S. L. (2002), "The medical treatment of depression, 1991-1996: productive inefficiency, expected outcome variations, and price indexes," *Journal of Health Economics*, 21, 373–396.

Buntin, M. B. and Zaslavsky, A. M. (2004), "Too much ado about two-part models and transformation?: Comparing methods of modeling Medicare expenditures," *Journal of Health Economics*, 23, 525–542.

Chai, H. S. and Bailey, K. R. (2008), "Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero," *Statistics in Medicine*, 27, 3643–3655.

Cooper, N. J., Lambert, P. C., Abrams, K. R., and Sutton, A. J. (2007), "Predicting costs over time using Bayesian Markov chain Monte Carlo methods: an application to early inflammatory polyarthritis," *Health Economics*, 16, 37–56.

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

Diehr, P., Yanez, D., Ash, A., Hornbrook, M., and Lin, D. (1999), "Methods for analyzing health care utilization and costs," *Annual Review of Public Health*, 20, 125–144.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002), *Analysis of Longitudinal Data*, Oxford University Press.

Duan, N., Manning, Jr., W. G., Morris, C. N., and Newhouse, J. P. (1983), "A comparison of alternative models for the demand of medical care," *Journal of Business and Economic Statistics*, 1, 115–126.

Finkelstein, E. A., Trogdon, J. G., Cohen, J. W., and Dietz, W. (2009), "Annual medical spending attributable to obesity: payer-and service-specific estimates," *Health Affairs*, 28, w822–w831.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012), *Applied Longitudinal Analysis*, John Wiley & Sons.

Ghosh, P. and Albert, P. S. (2009), "A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial," *Computational Statistics & Data Analysis*, 53, 699–706.

Hall, D. B. and Zhang, Z. (2004), "Marginal models for zero inflated clustered data," *Statistical Modelling*, 4, 161–180.

Hernan, M. A. and Robins, J. M. (2014), "Causal Infererence," Retrieved February 21, 2014 from Harvard University website.

Kahwati, L. C., Lance, T. X., Jones, K. R., and Kinsinger, L. S. (2011), "RE-AIM evaluation of the Veterans Health Administration's MOVE! weight management program," *Translational Behavioral Medicine*, 1, 551–560.

Kauermann, G. and Carroll, R. J. (2001), "A note on the efficiency of sandwich covariance matrix estimation," *Journal of the American Statistical Association*, 96, 1387–1396.

Liu, L., Conaway, M. R., Knaus, W. A., and Bergin, J. D. (2008a), "A random effects four-part model, with application to correlated medical costs," *Computational Statistics & Data Analysis*, 52, 4458–4473.

Liu, L., Cowen, M. E., Strawderman, R. L., and Shih, Y.-C. T. (2010), "A flexible two-part random effects model for correlated medical costs," *Journal of Health Economics*, 29, 110–123.

Liu, L., Ma, J. Z., and Johnson, B. A. (2008b), "A multi-level two-part random effects model, with application to an alcohol-dependence study," *Statistics in Medicine*, 27, 3528–3539.

Liu, L., Strawderman, R. L., Johnson, B. A., and O'Quigley, J. M. (2012), "Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study," *Statistical Methods in Medical Research*.

Long, D. L., Preisser, J. S., Herring, A. H., and Golin, C. E. (2014), "A marginalized zero-inflated Poisson regression model with overall exposure effects," *Statistics in Medicine*, 33, 5151–5165.

Lu, S.-E., Lin, Y., and Shih, W.-C. J. (2004), "Analyzing excessive no changes in clinical trials with clustered data," *Biometrics*, 60, 257–267.

Maciejewski, M. L., Bryson, C., Perkins, M., Blough, D., Cunningham, F., Fortney, J., Krein, S., Stroupe, K., Sharp, N., and Liu, C. (2010a), "Increasing copayments and adherence to diabetes, hypertension and hyperlipidemic medications," *The American Journal of Managed Care*, 16, e20–e32.

Maciejewski, M. L., Liu, C.-F., Kavee, A. L., and Olsen, M. K. (2012a), "How price responsive is the demand for specialty care?" *Health Economics*, 21, 902–912.

Maciejewski, M. L., Livingston, E. H., Smith, V. A., Kahwati, L. C., Henderson, W. G., and Arterburn, D. E. (2012b), "Health expenditures among high-risk patients after gastric bypass and matched controls," *Archives of Surgery*, 147, 633–640.

Maciejewski, M. L., Smith, V. A., Livingston, E. H., Kavee, A. L., Kahwati, L. C., Henderson, W. G., and Arterburn, D. E. (2010b), "Health care utilization and expenditure changes associated with bariatric surgery," *Medical Care*, 48, 989–998.

Madden, C. W., Mackay, B. P., Skillman, S. M., Ciol, M., and Diehr, P. K. (2000), "Risk adjusting capitation: applications in employed and disabled populations," *Health Care Management Science*, 3, 101–109.

Manning, W. G., Basu, A., and Mullahy, J. (2005), "Generalized modeling approaches to risk adjustment of skewed outcomes data," *Journal of Health Economics*, 24, 465–488.

Manning, W. G., Morris, C. N., Newhouse, J. P., Orr, L. L., Duan, N., Keeler, E., Leibowitz, A., Marquis, K., Marquis, M., and Phelps, C. (1981), "A two-part model of the demand for medical care: preliminary results from the health insurance study," *Health, Economics, and Health Economics*, 103–123.

Manning, W. G. and Mullahy, J. (2001), "Estimating log models: to transform or not to transform?" *Journal of Health Economics*, 20, 461–494.

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

Neelon, B., O'Malley, A. J., and Normand, S.-L. T. (2011), "A Bayesian two-part latent class model for longitudinal medical expenditure data: assessing the impact of mental health and substance abuse parity," *Biometrics*, 67, 280–289.

Olsen, M. K. and Schafer, J. L. (2001), "A two-part random-effects model for semicontinuous longitudinal data," *Journal of the American Statistical Association*, 96, 730–745.

Park, R. E. (1966), "Estimation with heteroscedastic error terms," *Econometrica*, 34, 888.

Parsons (2001), *Reducing bias in a propensity score matched-pair sample using greedy matching techniques*, Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Royall, R. M. (1986), "Model robust confidence intervals using maximum likelihood estimators," *International Statistical Review*, 221–226.

Smith, V. A., Preisser, J. S., Neelon, B., and Maciejewski, M. L. (2014), "A marginalized two-part model for semicontinuous data," *Statistics in Medicine*, 33, 4891–4903.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.

Su, L., Tom, B. D., and Farewell, V. T. (2011), "A likelihood-based two-part marginal model for longitudinal semi-continuous data," *Statistical methods in medical research*.

Su, L., Tom, B. D. M., and Farewell, V. T. (2009), "Bias in 2-part mixed models for longitudinal semicontinuous data," *Biostatistics*, 10, 374–389.

Tom, B. D., Su, L., and Farewell, V. T. (2013), "A corrected formulation for marginal inference derived from two-part mixed models for longitudinal semi-continuous data," *Statistical Methods in Medical Research*.

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002), "Analysis of repeated measures data with clumping at zero," *Statistical Methods in Medical Research*, 11, 341–355.

Xie, H., McHugo, G., Sengupta, A., Clark, R., and Drake, R. (2004), "A method for analyzing longitudinal outcomes with many zeros," *Mental Health Services Research*, 6, 239–246.

Zhang, M., Strawderman, R. L., Cowen, M. E., and Wells, M. T. (2006), "Bayesian inference for a two-part hierarchical model: an application to profiling providers in managed health care," *Journal of the American Statistical Association*, 101, 934–945.