Check for updates

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2018, Vol. 27(10) 3039–3061
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280217690770
journals.sagepub.com/home/smm
SAGE

Estimating marginal and incremental effects in the analysis of medical expenditure panel data using marginalized two-part random-effects generalized Gamma models: Evidence from China healthcare cost data

Bo Zhang,[1] Wei Liu[2] and Yingyao Hu[3]

Abstract

Conditional two-part random-effects models have been proposed for the analysis of healthcare cost panel data that contain both zero costs from the non-users of healthcare facilities and positive costs from the users. These models have been extended to accommodate more flexible data structures when using the generalized Gamma distribution to model the positive healthcare expenditures. However, a major drawback with the extended model, which is inherited from the conditional models, is that it is fairly difficult to make direct marginal inference with respect to overall healthcare costs that includes both zeros and non-zeros, or even on positive healthcare costs. In this article, we first propose two types of marginalized two-part random-effects generalized Gamma models (m2RGGMs): Type I m2RGGMs for the inference on positive healthcare costs and Type II m2RGGMs for the inference on overall healthcare costs. Then, the concepts of marginal effect and incremental effect of a covariate on overall and positive healthcare costs are introduced, and estimation of these effects is carefully discussed. Especially, we derive the variance estimates of these effects by following the delta methods and Taylor series approximations for the purpose of making marginal inference. Parameter estimates of Type I and Type II m2RGGMs are obtained through maximum likelihood estimation. An empirical analysis of longitudinal healthcare costs collected in the China Health and Nutrition Survey is conducted using the proposed methodologies.

Keywords

Panel data, random-effects models, marginal effects, healthcare expenditures, two-part models, marginal models

1 Introduction

Healthcare cost data collected in health service research and health economics studies exhibit the prominent features that the non-negative expenditure data (i) are substantially right-skewed with a heavy right tail, (ii) have excess zeros from the non-users of health services, and (iii) have heteroscedastic variance with respect to at least one of the covariates. Quantitative methods have been proposed in the literature to account for these features in analyzing cross-sectional healthcare expenditure data.[1–3] Manning et al.,[1] as well as Mullahy[2] and Blough et al.,[3] recognized the semicontinuous nature of health cost data (i.e., the data comprise a substantial portion of zero observations in healthcare costs from non-users and positive costs from health service users) and developed the two-part models with a logistic or probit model for

[1]Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD, USA
[2]Department of Mathematics, Harbin Institute of Technology, Harbin, P.R. China
[3]Department of Economics, Johns Hopkins University, Baltimore, MD, USA

Corresponding authors:
Bo Zhang, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA.
Email: Bo.Zhang@fda.hhs.gov
Wei Liu, Department of Mathematics, Harbin Institute of Technology, 92 West Dazhi Street, Nangang District, Harbin, Heilongjiang 150001, P.R. China.
Email: liuwhit@hit.edu.cn

the probability of observing a positive cost in part (I) and a generalized linear regression model for the observed positive healthcare costs in part (II). However, a large number of empirical studies are designed to repeatedly collect healthcare costs from the same cross section of individuals over time. As a result, many rich healthcare cost panel data, such as the Medical Expenditure Panel Survey data (https://meps.ahrq.gov) and the German Socioeconomic Panel data (https://www.diw.de/en/soep), have been developed in the area of health economics. The two-part models have been extended in order to accommodate the repeated measures in healthcare cost panel data, but the extended models still have significant limitations in term of making inference on marginal effects of covariates.

Quantitative methods that incorporate latent random effects into the analysis of healthcost panel data initially emerged in the work reported by Olsen and Schafer[4] and Tooze et al.[5] Olsen and Schafer[4] extended the two-part regression models for cross-sectional data to the settings of panel data, by introducing unobserved random coefficients into both parts of the models. To obtain maximum likelihood estimates, Olsen and Schafer[4] created an approximate Fisher scoring algorithm, in which a sixth-order multivariate Laplace approximation was employed to numerically evaluate the integrals in marginal likelihood. Tooze et al.[5] introduced a similar two-part random-effects model as in Olsen and Schafer,[4] assuming a lognormal model for the observed positive healthcare costs in part (II) instead of a normal model in Olsen and Schafer.[4] Tooze et al.[5] also distinguished their work from Olsen and Schafer[4] in computational strategy as they maximized the marginal likelihood using quasi-Newton optimization, in which the integrals were approximated by an adaptive Gaussian quadrature. The rationale behind these two-part random-effects models is that the random effects from part (I) and part (II) are assumed to be correlated so that the probability of observing a positive cost in part (I) is correlated to the amount of the positive cost. Of course, the random effects are able to characterize individual variation in each model component as usual. The two-part random-effects models typically employ a logistic or probit random-effects model in part (I) for the binary outcomes of presence or absence of healthcare expenditures. However, in part (II), there has been some debate on whether or not the use of a transformation on healthcare cost outcomes is preferred over the sophisticated parametric models. This debate applies to the empirical analysis of both cross-sectional and panel healthcare expenditure data. At the beginning, Duan et al.[6] proposed the "smearing" method, promoting the approach of estimating the mean of a untransformed healthcare cost outcomes after fitting a linear regression model with a nonparametric error term to the transformed outcomes. Yet, Mullahy,[2] Zhou et al.,[7] and Zhou et al.[8] showed that this "transformation and re-transformation" technique had its embedded drawback as it cannot stabilize the variance and thus heteroscedasticity may be omitted. Especially, Mullahy[2] described the circumstances where the conventional two-part models with homoscedastic transformation failed to provide consistent inference on marginal effects of covariates. To address these concerns, Manning et al.[9] proposed to use the three-parameter generalized Gamma distribution (GGD) for modeling the positive healthcare expenditures. GGD is particularly superior to others because (i) it avoids the troubles brought by conducting transformation and then re-transformation, (ii) it includes the standard Gamma, inverse Gamma, Weibull, exceptional, and lognormal distributions as its special cases, (iii) it provides much more flexibility in the circumstances where none of above distributions adequately fit the data, and (iv) it allows for the existence of substantial heteroscedasticity. Taking the advantages of the three-parameter GGD in Manning et al.,[9] Liu et al.[10] constructed a flexible two-part random-effects model for clustered medical expenditures, in which the lognormal random-effects model developed by Tooze et al.[5] in Part (II) was extended to be a generalized Gamma regression model with latent random effects and with a scale parameter that was allowed to depend on covariates.

This article is devoted to developing two types of marginalized two-part random-effects generalized Gamma models (m2RGGMs) that can achieve the goal of making direct inference on the marginal effects of covariates[11,12] in both Part (I) and Part (II) of the two-part models and making inference on the marginal effects of covariates with respect to both positive and overall healthcare costs. The fact is that the two-part random-effects models discussed in Olsen and Schafer,[4] Tooze et al.,[5] and Duan et al.[6] are conditional models. Liu et al.[10] encountered difficulties when formulating the strategy of conducting marginal inference in the context of the conditional two-part random-effects generalized Gamma models. This is because the conditional models have their inherent drawback in that the conditional models target at subject-specific characteristics instead of population-average interpretations and therefore a tedious differentiation procedure is inevitable in deriving estimates for marginal effects. It will be even harder to obtain variances of the estimates of marginal effects if statistical inference on the marginal effects is required for the conditional models. Yet, the primary objective of health service research and health economics studies is usually to make marginal inference on the treatment effect of a policy intervention or the partial effects of some covariates with respect to the expected amount of overall healthcare costs. In such circumstances, a marginal mean model is much more convenient, especially for the analysis of panel data. In this article, we describe how the correlation among healthcare costs over time can be characterized by embedding the marginal mean structure within a complete multivariate probability model based on random effects dependence assumption. This is achieved in both Part (I) and Part (II) of the two-part models, and eventually completes the construction of m2RGGMs in which Part (II) is built upon a generalized Gamma regression. There are two types of marginalized models proposed in this article: Type I m2RGGMs and Type II m2RGGMs. In the Type I m2RGGMs, the mean (or the first moment) regression parameters represent the change in expected positive healthcare expenditure, whereas in the Type II m2RGGMs, the regression parameters represent the change in the overall mean of healthcare expenditures including both zeros and non-zeros. In both types of models, correlations among healthcare expenditure responses given the covariates, even if reasonably attributed to shared unobservable latent variables, is accounted for by a separate dependence model in both Part (I) and Part (II). The motivation to further propose the Type II m2RGGM over the Type I m2RGGM is well supported by the empirical examples in the literature. Liu et al.[10] reported an empirical example, in which the pharmacy cost data collected from mid-western US managed care organization on 56,245 adult patients served by 239 primary care physicians were analyzed. The primary interest of the empirical analysis in Liu et al.[10] lied in identifying the patient-level factors that may affect the overall pharmaceutical expenditures that included both positive costs and zero costs. Huang and Gan[13] investigated the impact of Urban Employee Basic Medical Insurance (UEBMI) and other individual observable characteristics on healthcare expenditures among urban residents in China. Huang and Gan[13] also focused on the overall healthcare expenditures, instead of only investigating the positive costs. There are several advantages of the marginal approach in the proposed Type I and Type II m2RGGMs. First, the interpretation of the regression parameters on the positive and overall healthcare cost mean is invariant with respect to specification of the random-effects dependence model. Analysis with the marginal mean regression model but different association models has exactly the same target of marginal inference. Second, by introducing the marginally specified models in both Part (I) and Part (II) of the two-part models, we allow a choice as to whether the marginal mean structure or the conditional mean structure is the focus of modeling when using a latent random effects formulation. Third, by integrating both conditional and marginal models together into one two-part model, m2RGGMs allow making both conditional (subject-specific) and marginal (population-average) inferences simultaneously. Lastly, the m2RGGMs inherit all attributes of the generalized Gamma model in Manning et al.[9] to allow for more flexible and more adequate model fitting than the lognormal models and simultaneously allow for variance heteroscedasticity.

The scientific contributions of this article are twofold. Besides the development of m2RGGMs, this article is also devoted to deriving the estimates, as well as the variance estimates, of marginal effects of a covariate with respect to both overall and positive healthcare costs when the m2RGGMs are used. Marginal effect, or partial effect, is a basic concept and often the quantity of interest in econometrics.[14] Two variants of the marginal effect, elasticity and semi-elasticity, are also important quantities for the interpretation of empirical analysis. Basu and Rathouz[15] defined incremental effect, an analogous parameter of marginal effect, for discrete covariates. In this article, we derive the marginal effects, including the incremental effect, elasticity and semi-elasticity, for the proposed m2RGGMs and provide explicit formulas for the estimates of these effects. Note that there are two classes of marginal effects introduced in this article: (conventional) marginal effects and average marginal effects. Marginal effects refer to the conventional marginal effects whose estimates depend on both parameter estimates and specific (or subject-specific) covariate values, whereas average marginal effects are the population average of the conventional marginal effects. Here, we actually provide the estimates of both classes of marginal effects for the m2RGGM. More importantly, we use the delta methods and Taylor series approximations to obtain the variance estimates of these marginal effects and average marginal effects for the m2RGGMs. This is essential to making direct marginal inferences with these quantities when the m2RGGMs are employed. To demonstrate the capabilities of the proposed methodologies in empirical analysis, a set of longitudinal healthcare cost data collected in the China Health and Nutrition Survey (CHNS) is analyzed.

2 Modeling framework

2.1 The generalized Gamma distribution

GGD as a generalization of the standard two-parameter Gamma distribution is a continuous probability distribution with one-scale parameter and two-shape parameters. For a non-negative response variable y, the probability density function for the GGD is parameterized as a function of the scale parameter μ and the shape parameters σ and κ[9]

$$f_{GG}(y;\mu,\sigma,\kappa)=\frac{\gamma^{\gamma}}{\sigma y\sqrt{\gamma}\Gamma(\gamma)}\exp(z\sqrt{\gamma}-u),\quad y\geq 0 \tag{1}$$

where $\gamma = |\kappa|^{-2}$, $z = \text{sign}(\kappa)\{\log(y) - \mu\}/\sigma$ and $u = \gamma \exp(|\kappa|z)$. The formulation of density (1) is nicely presented by Manning et al.[9] for modeling skewed cross-sectional medical cost data. GGD is very appealing for modeling positive medical expenditure data, as it includes Gamma, inverse Gamma, Weibull, exponential, and log-normal distributions as its special cases. When $\kappa = 1$, the GGD becomes the Weibull distribution and the density (1) reduces to $\frac{1}{\sigma y}\exp\left[\frac{\log(y)-\mu}{\sigma} - \exp\left\{\frac{\log(y)-\mu}{\sigma}\right\}\right]$. When $\kappa \to 0$, the GGD converges to the log-normal distribution with a density of $\frac{1}{\sigma y\sqrt{2\pi}}\exp\left[-\frac{\{\log(y)-\mu\}^2}{2\sigma^2}\right]$. If $\sigma = \kappa > 0$, the GGD degenerates to the Gamma distribution and the density (1) reduces to $\frac{1}{\sigma^{2\sigma^{-2}} y\Gamma(\sigma^{-2})}\exp\left[\frac{\log(y)-\mu-y\exp(-\mu)}{\sigma^2}\right]$. If $\sigma = -\kappa > 0$, the GGD degenerates to the inverse Gamma distribution and the density (1) reduces to $\frac{1}{\sigma^{2\sigma^{-2}} y\Gamma(\sigma^{-2})}\exp\left[\frac{\mu-\log(y)-\exp(\mu)/y}{\sigma^2}\right]$. When $\kappa = \sigma = 1$, the GGD becomes the exponential distribution with a density of $\exp\left[-\frac{y}{\exp(\mu)} - \mu\right]$.

The generalized Gamma regression model is constructed through the parameter μ that is replaced by $\mu(x) = x'\beta$, where x is the vector of the covariates including an intercept, and β is the vector of coefficients to be estimated. The generalized Gamma regression model can be extended to allow for heteroscedasticity by parameterizing $\log\{\sigma(x)\} = x'\alpha$. For the GGD, the rth moment is given by

$$E(y^r) = \exp(r\mu)\frac{\kappa^{2r\sigma/\kappa}\Gamma(1/\kappa^2 + r\sigma/\kappa)}{\Gamma(1/\kappa^2)} \tag{2}$$

From (2), the expected value and variance of y conditional on x in the heteroscedastic generalized Gamma regression model can be derived as

$$E(y|x) = \exp\{\mu(x)\}\frac{\kappa^{2\sigma(x)/\kappa}\Gamma(1/\kappa^2 + \sigma(x)/\kappa)}{\Gamma(1/\kappa^2)} \tag{3}$$

and

$$\text{var}(y|x) = \exp\{2\mu(x)\}\kappa^{4\sigma(x)/\kappa}\left[\left\{\frac{\Gamma(1/\kappa^2 + 2\sigma(x)/\kappa)}{\Gamma(1/\kappa^2)}\right\} - \left\{\frac{\Gamma(1/\kappa^2 + \sigma(x)/\kappa)}{\Gamma(1/\kappa^2)}\right\}^2\right]$$

respectively.

2.2 Conditional two-part random-effects models

In this subsection, we reiterate the conditional model proposed in Liu et al.[10] with minor modification to conform healthcare cost panel data. Suppose a set of healthcare cost panel data is collected repeatedly on the same cross section of individuals over time. Denote Y_{it}, a semicontinuous (either zero or positive) response, as the amount of the healthcare cost of the ith individual measured at recorded time t and denote X_{it} as a K-dimensional covariate vector, in which $t = 1, 2, \ldots, T$ and $i = 1, 2, \ldots, N$ indexes the N individuals in the study. The conditional two-part random-effects generalized Gamma model (c2RGGM) introduced in Liu et al.[10] consists of two parts. In Part (I), as in standard two-part random-effects models, it is assumed that, conditioning on covariates X_{it} and the vector of unobserved random effects $a_i = (a_{i1}, \ldots, a_{iT})'$, $Y_{i1}, \ldots, Y_{iT}$ are independent and the probability of observing a positive cost in Y_{it} is formulated as

$$\hbar_{c_1}[P(Y_{it} > 0|X_{it}, a_{it})] = X'_{it}\alpha_c + a_{it} \tag{4}$$

in which $\hbar_{c_1}(\cdot)$ denotes a link function (such as a probit link function or a logit link function) and α_c is a vector of regression coefficients. A single scalar random effect a_{it} is used in (4), but this representation is more general because it includes several common models as its special cases: the random-intercept model corresponds to assuming $a_{it} = a_{i0}$ for $t = 1, 2, \ldots, T$; the mixed-effects model corresponds to assuming $a_{it} = Z'_{it}a^*_i$, where a^*_i is a q-dimensional random effect vector and the vector Z_{it} includes the covariates pertaining to the random effects; among others. In Part (II), it is assumed that, conditional on covariates X_{it} and the vector of unobserved random effects $b_i = (b_{i1}, \ldots, b_{iT})'$ and given $Y_{it} > 0$ for $t = 1, 2, \ldots, T$, $Y_{i1}, \ldots, Y_{iT}$ are independent and Y_{it} follows GGD with a density of $f_{GG}(y_{it}; \kappa, \mu_{it}, \sigma_{it})$. The effects of covariates are connected to the parameters through

$$(\mu_{it} | Y_{it} > 0, X_{it}, b_{it}) = X'_{it}\beta_c + b_{it} \tag{5}$$

and $\log(\sigma_{it} | Y_{it} > 0, X_{it}) = X'_{it}\delta_c$, in which β_c and δ_c are two vectors of regression coefficients. The scalar random effect b_{it} is similarly formulated as the a_{it} in (4): assuming $b_{it} = b_{i0}$ for random-intercept models, or assuming $b_{it} = Z'_{it}b^*_i$ for mixed-effects models. The random effects a_{it} in (4) and b_{it} in (5) are assumed to be correlated and follow a multivariate normal distribution. The format and dimension of the multivariate normal distribution depend on the specifications of a_{it} and b_{it}. If (4) and (5) are two random-intercept models or equivalently $a_{it} = a_{i0}$ and $b_{it} = b_{i0}$, then (a_{i0}, b_{i0}) follows a bivariate normal distribution. If (4) and (5) are two mixed-effects models with $a_{it} = Z'_{it}a^*_i$ and $b_{it} = Z'_{it}b^*_i$, then it can be assumed that (a^*_i, b^*_i) follows a $(2q)$-dimensional multivariate normal distribution $N_{2q}(0, \tilde{\Sigma}_i)$ and consequently (a_{it}, b_{it}) follows $N_2(0, (Z'_{it}, Z'_{it})\tilde{\Sigma}_i(Z'_{it}, Z'_{it})')$. In both cases, the odds of having a non-zero cost and the amount of positive cost for an individual are associated through the correlation between a_{it} in (4) and b_{it} in (5). If the correlation coefficients in $\text{cov}(a_i, b_i)$ are zero, then the two parts of the model are separated, indicating that the presence or absence of healthcare cost at one occasion has no influence on the amount, if any, at this or other occasions. In practice, however, this is frequently not true and therefore the random effects from two parts are usually correlated. In c2RGGMs, intercepts and slopes of the covariates in X_{it} for either Part (I) and Part (II) may be fixed or random. Additional static or time-varying covariates also may be included in either one. The same set of covariates may appear as it is presented now in (4) and (5), but this is not required. If it is needed, different covariates can be used by simply fixing the corresponding fixed or random effects at zero. Healthcare expenditure responses Y_{it}'s in c2RGGMs need not be recorded at the same set of time points for all individuals, and the data may be unbalanced by design or have ignorably missing values. In c2RGGMs, the log-linear dependence of σ_{it} on covariates X_{it} allows for heteroscedasticity, and hypothesis testing procedures can formally test for the existence of heteroscedasticity in c2RGGMs.

2.3 Type I marginalized two-part random-effects models

The c2RGGMs introduced in Section 2.2 cannot directly make marginal inference and thus the interpretation of α_c and β_c can be particularly difficult. In this subsection, we propose to modify the c2RGGMs by augmenting the marginal mean model, so that a direct marginal inference can be achieved in two-part models for healthcare cost panel data.

In Part (I), a marginal model for the probability that the ith individual has a positive healthcare cost at the recorded time t is given as

$$\hbar_{m_1}[P(Y_{it} > 0 | X_{it})] = X'_{it}\alpha_m \tag{6}$$

in which $\hbar_{m_1}(\cdot)$ denotes the link function and α_m is a vector of marginal regression coefficients. However, (6) only identifies the marginal mean (first moment) of the complete multivariate distribution of $Y_i = (Y_{i1}, \ldots, Y_{iT})'$. In order to complete model specification, the dependence among repeated measures of healthcare costs is required. As a result, a conditional model that characterizes the dependence among healthcare costs Y_{it}'s is further specified as

$$\hbar_{c_1}[P(Y_{it} > 0 | X_{it}, a_{it})] = \Delta_{it}(X'_{it}\alpha_m, \text{var}(a_{it})) + a_{it} \tag{7}$$

In (7), $\hbar_{c_1}(\cdot)$ denotes the link function and a_{it} is a random effect as described in Section 2.2. The link function in (6) and (7) can in general be different, although in practice we usually choose them to be identical. It is assumed that, given the latent random vector $a_i = (a_{i1}, \ldots, a_{iT})'$, $Y_{i1}, \ldots, Y_{iT}$ are independent and that in general $a_i \sim N(0, A_i(\theta_A))$, in which the covariance matrix A_i is a function of random effect covariates Z_{it}'s and the vector θ_A that contains variance-covariance parameters. In (7), the parameter $\Delta_{it} = \Delta_{it}(X'_{it}\alpha_m, \text{var}(a_{it}))$ is a function of marginal mean predictor $X'_{it}\alpha_m$, the random effect variance $\text{var}(a_{it})$, and possibly other parameters. Equations (6) and (7) together constitute the marginally specified model for the probability of observing a positive healthcare cost in Y_{it}[11,12], in which (6) captures the systematic variation in the mean that is due to X_{it} and (7) provides measures of random variation both across individuals and over time. By this way, the marginalized models (6) and (7) separate the model for systematic variation from the model for random variation. From (6), we have

$$P(Y_{it} > 0|X_{it}) = \hbar_{m_1}^{-1}(X'_{it}\alpha_m)$$

If $f_{a_{it}}(\cdot)$ denotes the distribution of a_{it} and $\phi(\cdot)$ denotes the probability density function of a standard normal distribution, (7) indicates that

$$P(Y_{it} > 0|X_{it}) = \int P(Y_{it} > 0|X_{it}, a_{it})f_{a_{it}}(a_{it})\mathrm{d}a_{it} = \int \hbar_{c_1}^{-1}[\Delta_{it} + \mathrm{var}(a_{it})^{1/2}s]\phi(s)\mathrm{d}s$$

Therefore, Δ_{it} can be obtained as the solution to the convolution equation

$$\hbar_{m_1}^{-1}(X'_{it}\alpha_m) = \int \hbar_{c_1}^{-1}[\Delta_{it} + \mathrm{var}(a_{it})^{1/2}s]\phi(s)\mathrm{d}s \tag{8}$$

This equation can be solved for Δ_{it} using numerical integration combined with Newton-Raphson or quasi-Newton algorithm, but a close-form solution may exist in some cases.

In Part (II), we propose the following marginally specified model for the mean of the positive healthcare costs

$$\hbar_{m_2}[E(Y_{it}|Y_{it} > 0, X_{it})] = X'_{it}\beta_m \tag{9}$$

in which $\hbar_{m_2}(\cdot)$ is a link function and β_m is the vector of marginal regression coefficients for the positive healthcare costs. The rationale of (9) is to provide direct marginal inference on the mean of the positive healthcare costs while using the two-part models for healthcare cost panel data. In order to characterize the dependence among repeated measurements on positive costs, we assume that, conditioning on covariates X_{it}, random effect b_{it} and $Y_{it} > 0$, Y_{it} follows GGD with a density of $f_{GG}(y_{it}; \kappa, \mu_{it}, \sigma_{it})$. Furthermore, a conditional model is constructed for μ_{it} as

$$(\mu_{it}|Y_{it} > 0, X_{it}, b_{it}) = \Lambda_{it}(X'_{it}\beta_m, X'_{it}\delta_m, \mathrm{var}(b_{it})) + b_{it} \tag{10}$$

and $\log(\sigma_{it}|Y_{it} > 0, X_{it}) = X'_{it}\delta_m$, in which b_{it} is a random effect as described in Section 2.2, δ_m is the vector of regression coefficients for σ_{it}, and the parameter $\Lambda_{it} = \Lambda_{it}(X'_{it}\beta_m, X'_{it}\delta_m, \mathrm{var}(b_{it}))$ is a function of linear predictors $X'_{it}\beta_m$, $X'_{it}\delta_m$, the random effect variance $\mathrm{var}(b_{it})$, and possibly other parameters. It is assumed that $Y_{i1}, \ldots, Y_{iT}$ are independent given the unobserved random vector $b_i = (b_{i1}, \ldots, b_{iT})'$ and that $b_i \sim N(0, B_i(\theta_B))$, in which the covariance matrix B_i is a function of Z_{it}'s and the variance-covariance parameter vector θ_B. From (9), we have

$$E(Y_{it}|Y_{it} > 0, X_{it}) = \hbar_{m_2}^{-1}(X'_{it}\beta_m)$$

It can be derived from (10) that

$$\begin{aligned} E(Y_{it}|Y_{it} > 0, X_{it}) &= \int E(Y_{it}|Y_{it} > 0, X_{it}, b_{it})f_{b_{it}}(b_{it})\mathrm{d}b_{it} \\ &= \int \exp[\Lambda_{it} + \mathrm{var}(b_{it})^{1/2}s]\frac{\kappa^{2\exp(X'_{it}\delta_m)/\kappa}\Gamma(1/\kappa^2 + \exp(X'_{it}\delta_m)/\kappa)}{\Gamma(1/\kappa^2)}\phi(s)\mathrm{d}s \\ &= \frac{\kappa^{2\exp(X'_{it}\delta_m)/\kappa}\Gamma(1/\kappa^2 + \exp(X'_{it}\delta_m)/\kappa)}{\Gamma(1/\kappa^2)}\exp\left[\Lambda_{it} + \frac{\mathrm{var}(b_{it})}{2}\right] \end{aligned}$$

Therefore, (9) and (10) lead to

$$\Lambda_{it} = \log\left[\frac{\hbar_{m_2}^{-1}(X'_{it}\beta_m)\Gamma(1/\kappa^2)}{\kappa^{2\exp(X'_{it}\delta_m)/\kappa}\Gamma(1/\kappa^2 + \exp(X'_{it}\delta_m)/\kappa)}\right] - \frac{\mathrm{var}(b_{it})}{2} \tag{11}$$

Models (6), (7), (9), and (10) together constitute a m2RGGM that emphasizes the inference of the marginal effects of covariates on positive costs. Here we name it as "Type I m2RGGM", in contrast to the "Type II m2RGGM" introduced in Section 2.4. As in Section 2.2, the random effects a_{it} in (7) and b_{it} in (10) are assumed to be correlated and follow a multivariate normal distribution so that the two parts are correlated; that is, we assume generally $(a_{it}, b_{it})' \sim N_2(0, \Sigma(\theta_A, \theta_B, \theta_{AB}))$, in which θ_{AB} is a vector that contains those additional unknown parameters in Σ other than the ones in θ_A and θ_B. Note that there is a critical distinction between marginal parameters α_m and β_m in m2RGGMs and the conditional parameters α_c and β_c in c2RGGMs. The conditional regression coefficients contrasts the expected healthcare cost response in different covariate values under the condition that the values of the latent random effects are equivalent, whereas the marginal coefficients do not control for the unobserved random effects when characterizing the probabilities of positive costs or the mean of positive costs. A marginal treatment effect directly compares the mean of treatment group to the mean of control group, while a conditional treatment effect compares these means assuming the latent random effects are equal.

2.4 Type II marginalized two-part random-effects models

In health economics and health service research, the inference on the overall mean of the healthcare expenditures that include both zero and non-zero records may be much more influential than the inference on the marginal effects of covariates on positive costs only. Apparently, the proposed Type I m2RGGMs in Section 2.3 are not able to achieve the goal of providing direct marginal inference with respect to the marginal mean of overall healthcare costs (including both zero and positive costs) over time. In order to accomplish this goal, we propose to use the following marginally specified model to replace (9) in Type I m2RGGMs

$$\hbar_{m_3}[E(Y_{it}|X_{it})] = X'_{it}\beta_m \tag{12}$$

in which $\hbar_{m_3}(\cdot)$ is a link function and β_m is the vector of marginal regression coefficients for overall healthcare costs. The rationale of this proposal is to promote direct marginal inference on the overall healthcare costs, while still using a similar parametric specification in (10) for Part (II) of Type I m2RGGMs. As in (10), we still assume that, conditioning on covariates X_{it}, random effect b_{it} and $Y_{it} > 0$, Y_{it} follows GGD with a density of $f_{GG}(y_{it}; \kappa, \mu_{it}, \sigma_{it})$. In addition, a conditional model is constructed for μ_{it} as

$$(\mu_{it}|Y_{it} > 0, X_{it}, b_{it}) = \Lambda_{it}(X'_{it}\alpha_m, X'_{it}\beta_m, X'_{it}\delta_m, \mathrm{var}(b_{it})) + b_{it} \tag{13}$$

and $\log(\sigma_{it}|Y_{it} > 0, X_{it}) = X'_{it}\delta_m$, in which the parameter $\Lambda_{it} = \Lambda_{it}(X'_{it}\alpha_m, X'_{it}\beta_m, X'_{it}\delta_m, \mathrm{var}(b_{it}))$ is a function of linear predictors $X'_{it}\alpha_m$, $X'_{it}\beta_m$, $X'_{it}\delta_m$, the random effect variance $\mathrm{var}(b_{it})$, and possibly other parameters. Models (12) and (13) together constitute the marginally specified model for Part (II) of Type II m2RGGMs. Part (I) of Type II m2RGGMs remain identical to that of Type I m2RGGMs. As in Section 2.3, with the assumption that the random effects a_{it} in (7) and b_{it} in (10) jointly follow a multivariate normal distribution. Models (6), (7), (12), and (13) together constitute the Type II m2RGGM, which emphasizes the inference of the marginal effects of covariates on the overall mean of the panel healthcare costs. From (12), we have

$$E(Y_{it}|X_{it}) = \hbar^{-1}_{m_3}(X'_{it}\beta_m)$$

It can be derived from (6) and (13) that

$$\begin{aligned} E(Y_{it}|X_{it}) &= P(Y_{it} > 0|X_{it})E(Y_{it}|Y_{it} > 0, X_{it}) \\ &= \hbar^{-1}_{m_1}(X'_{it}\alpha_m)\frac{\kappa^{2\exp(X'_{it}\delta_m)/\kappa}\Gamma(1/\kappa^2 + \exp(X'_{it}\delta_m)/\kappa)}{\Gamma(1/\kappa^2)}\exp\left[\Lambda_{it} + \frac{\mathrm{var}(b_{it})}{2}\right] \end{aligned}$$

Therefore, (10) and (12) lead to

$$\Lambda_{it} = \log\left[\frac{\hbar^{-1}_{m_3}(X'_{it}\beta_m)\Gamma(1/\kappa^2)}{\hbar^{-1}_{m_1}(X'_{it}\alpha_m)\kappa^{2\exp(X'_{it}\delta_m)/\kappa}\Gamma(1/\kappa^2 + \exp(X'_{it}\delta_m)/\kappa)}\right] - \frac{\mathrm{var}(b_{it})}{2} \tag{14}$$

Note that the proposed Type I m2RGGM in Section 2.3 cannot be subsumed under the Type II m2RGGM. The Type II m2RGGM specifies a marginal model (12) directly on the expected overall healthcare cost, whereas the counterpart in the Type I m2RGGM is the marginal model (9) that characterizes the marginal association between the covariates and the expected positive healthcare cost.

3 Marginal and incremental effects

In the panel studies on healthcare costs, the quantities of interest are often the marginal effects of changes in the covariates at a specific time point.[16–18] Let y_{it} be the semicontinuous response variable that represents the amount of the healthcare cost (either zero or positive cost) of the ith individual measured at the time t. Let $x_{it} = (x_{it1}, x_{it2}, \ldots, x_{itK})$ be a vector of K covariates. The marginal effect, or partial effect, of the kth covariate x_{itk} at the time t on the expected overall healthcare cost $E(y_{it}|x_{it})$ is defined as the partial derivative of $E(y_{it}|x_{it})$ with respect to covariate x_{itk}

$$\eta_k(x_{it}, \vartheta) = \frac{\partial E(y_{it}|x_{it})}{\partial x_{itk}} \tag{15}$$

in which ϑ denotes the vector that includes all unknown parameters (also see Section 4). The assumptions for (15) include that x_{itk} is continuous and that $E(y_{it}|x_{it})$ is differentiable with respect to x_{itk} for $k = 1, 2, \ldots, K$. The marginal effect allows us to quantify the marginal change in the expected overall healthcare cost when covariate x_{itk} is increased by a small amount while holding other covariates $x_{it,-k} = (x_{it1}, \ldots, x_{it,k-1}, x_{it,k+1}, \ldots, x_{itK})$ constant. Note that the marginal effect $\eta_k(x_{it}, \vartheta)$ is usually a function of both unknown parameters ϑ and covariates x_{it}, although it can be simplified when $E(y_{it}|x_{it})$ is a linear combination of the covariates. When x_{itk} is categorical representing multiple levels or experimental groups, the quantities of interest is usually the incremental effect[15] that is defined as

$$\pi_k(x_{it,-k}, \vartheta) = E(y_{it}|x_{itk} = l_2, x_{it,-k}) - E(y_{it}|x_{itk} = l_1, x_{it,-k})$$

in which l_1 and l_2 are two levels of x_{itk}. The incremental effect quantifies the difference in the expected overall healthcare cost $E(y_{it}|x_{it})$ at the two levels of x_{itk} while holding other covariates $x_{it,-k}$ constant. When x_{itk} is binary that takes values 1 and 0 to represent two experimental groups (e.g., treatment and control groups), the incremental effect is

$$\pi_k(x_{it,-k}, \vartheta) = E(y_{it}|x_{itk} = 1, x_{it,-k}) - E(y_{it}|x_{itk} = 0, x_{it,-k})$$

It is especially convenient to derive the marginal and incremental effects of a covariates on overall healthcare costs, when the Type II m2RGGMs are used. From (12), it can be directly derived for the Type II m2RGGMs that the marginal and incremental effects are

$$\eta_k(x_{it}, \vartheta) = \frac{\partial \hbar_{m_3}^{-1}(x_{it}'\beta_m)}{\partial(x_{it}'\beta_m)} \cdot \beta_{m,k} = \dot{\hbar}_{m_3}^{-1}(x_{it}'\beta_m) \cdot \beta_{m,k}$$

and

$$\pi_k(x_{it,-k}, \vartheta) = \hbar_{m_3}^{-1}[(x_{it1}, \ldots, x_{itk} = l_2, \ldots, x_{itK})\beta_m] - \hbar_{m_3}^{-1}[(x_{it1}, \ldots, x_{itk} = l_1, \ldots, x_{itK})\beta_m]$$

respectively. Here, $\beta_{m,k}$ denotes the regression coefficient for covariate x_{itk}. The dot on $\dot{\hbar}_{m_3}^{-1}$ indicates it is the derivative of $\hbar_{m_3}^{-1}(\cdot)$; that is, $\dot{\hbar}_{m_3}^{-1}(x) = \frac{\partial \hbar_{m_3}^{-1}(x)}{\partial x}$. If the link function $\hbar_{m_3}(\cdot)$ is a log link (i.e., $\hbar_{m_3}(\cdot) = \log(\cdot)$), then the marginal and incremental effects for the Type II m2RGGMs are

$$\eta_k(x_{it}, \vartheta) = \exp(x_{it}'\beta_m) \cdot \beta_{m,k}$$

and

$$\pi_k(x_{it,-k}, \vartheta) = \exp[(x_{it1}, \ldots, x_{itk} = l_2, \ldots, x_{itK})\beta_m] - \exp[(x_{it1}, \ldots, x_{itk} = l_1, \ldots, x_{itK})\beta_m]$$

respectively. In some empirical analysis of healthcare costs, two particular functions of the marginal effect are the quantities of primary interest: elasticity and semi-elasticity. In a panel study, the elasticity and semi-elasticity of the expected overall healthcare cost $E(y_{it}|x_{it})$ with respect to covariate x_{itk} are defined as

$$e_k(x_{it}, \vartheta) = \frac{\partial E(y_{it}|x_{it})/E(y_{it}|x_{it})}{\partial x_{itk}/x_{itk}} = \frac{\partial E(y_{it}|x_{it})}{\partial x_{itk}} \cdot \frac{x_{itk}}{E(y_{it}|x_{it})} = \frac{\partial \log E(y_{it}|x_{it})}{\partial \log x_{itk}} \tag{16}$$

and

$$s_k(x_{it}, \vartheta) = \frac{\partial E(y_{it}|x_{it})/E(y_{it}|x_{it})}{\partial x_{itk}} = \frac{\partial E(y_{it}|x_{it})}{\partial x_{itk}} \cdot \frac{1}{E(y_{it}|x_{it})} = \frac{\partial \log E(y_{it}|x_{it})}{\partial x_{itk}}$$

respectively. In (16), $x_{itk} > 0$ is assumed only when it is needed. The elasticity measures the percentage change in the expected overall healthcare cost $E(y_{it}|x_{it})$ to a percentage change in covariate x_{itk} at the time point t, while the semi-elasticity quantifies the percentage change in the expected overall healthcare cost $E(y_{it}|x_{it})$ when x_{itk} is increased by one unit. Specifically, the elasticity and semi-elasticity for the Type II m2RGGMs are

$$e_k(x_{it}, \vartheta) = \frac{\partial \log \hbar_{m_3}^{-1}(x'_{it}\beta_m)}{\partial \log x_{itk}} = \frac{\dot{\hbar}_{m_3}^{-1}(x'_{it}\beta_m)}{\hbar_{m_3}^{-1}(x'_{it}\beta_m)} x_{itk}\beta_{m,k}$$

and

$$s_k(x_{it}, \vartheta) = \frac{\partial \log \hbar_{m_3}^{-1}(x'_{it}\beta_m)}{\partial x_{itk}} = \frac{\dot{\hbar}_{m_3}^{-1}(x'_{it}\beta_m)}{\hbar_{m_3}^{-1}(x'_{it}\beta_m)} \beta_{m,k}$$

If the link function $\hbar_{m_3}(\cdot)$ is a log link, then the elasticity and semi-elasticity for the Type II m2RGGMs are

$$e_k(x_{it}, \vartheta) = x_{itk}\beta_{m,k}$$

and

$$s_k(x_{it}, \vartheta) = \beta_{m,k}$$

respectively. In contrast to the Type I m2RGGMs which will be discussed promptly, estimation on marginal effect $\eta_k(x_{it}, \vartheta)$, as well as incremental effect $\pi_k(x_{it,-k}, \vartheta)$, elasticity $e_k(x_{it}, \vartheta)$ and semi-elasticity $s_k(x_{it}, \vartheta)$ on the expected overall healthcare cost, is much more convenient in the Type II m2RGGMs due to the concise marginal expression in the models.

For the Type I m2RGGMs, the marginal effects of a covariate on the expected overall healthcare cost $E(y_{it}|x_{it})$ can also be derived. When the Type I m2RGGMs are used, we have from (6) and (9) that

$$E(Y_{it}|X_{it}) = P(Y_{it} > 0|X_{it})E(Y_{it}|Y_{it} > 0, X_{it}) = \hbar_{m_1}^{-1}(X'_{it}\alpha_m)\hbar_{m_2}^{-1}(X'_{it}\beta_m)$$

It implies that the marginal effect on the expected overall healthcare cost $E(y_{it}|x_{it})$ with respect to covariate x_{itk} is

$$\eta_k(x_{it}, \vartheta) = \dot{\hbar}_{m_1}^{-1}(x'_{it}\alpha_m)\hbar_{m_2}^{-1}(x'_{it}\beta_m) \cdot \alpha_{m,k} + \hbar_{m_1}^{-1}(x'_{it}\alpha_m)\dot{\hbar}_{m_2}^{-1}(x'_{it}\beta_m) \cdot \beta_{m,k}$$

in which $\alpha_{m,k}$ denotes the regression coefficient of covariate x_{itk} in (6). In addition, the elasticity and semi-elasticity of the expected overall healthcare cost $E(y_{it}|x_{it})$ when the Type I m2RGGMs are used can be derived as follows

$$\begin{aligned} e_k(x_{it}, \vartheta) &= \eta_k(x_{it}, \vartheta) \cdot \frac{x_{itk}}{E(y_{it}|x_{it})} \\ &= \frac{\dot{\hbar}_{m_1}^{-1}(x'_{it}\alpha_m)}{\hbar_{m_1}^{-1}(x'_{it}\alpha_m)} x_{itk}\alpha_{m,k} + \frac{\dot{\hbar}_{m_2}^{-1}(x'_{it}\beta_m)}{\hbar_{m_2}^{-1}(x'_{it}\beta_m)} x_{itk}\beta_{m,k} \end{aligned}$$

and

$$\begin{aligned} s_k(x_{it}, \vartheta) &= \eta_k(x_{it}, \vartheta) \cdot \frac{1}{E(y_{it}|x_{it})} \\ &= \frac{\dot{\hbar}_{m_1}^{-1}(x_{it}'\alpha_m)}{\hbar_{m_1}^{-1}(x_{it}'\alpha_m)}\alpha_{m,k} + \frac{\dot{\hbar}_{m_2}^{-1}(x_{it}'\beta_m)}{\hbar_{m_2}^{-1}(x_{it}'\beta_m)}\beta_{m,k} \end{aligned}$$

When x_{itk} is a categorical variable with multiple levels, the incremental effect on the expected overall healthcare cost $E(y_{it}|y_{it} > 0, x_{it})$ from level l_2 to level l_1 of covariate x_{itk} is

$$\begin{aligned} \pi_k(x_{it,-k}, \vartheta) = {} & \hbar_{m_1}^{-1}[(x_{it1}, \ldots, x_{itk} = l_2, \ldots, x_{itK})\alpha_m]\hbar_{m_2}^{-1}[(x_{it1}, \ldots, x_{itk} = l_2, \ldots, x_{itK})\beta_m] \\ & - \hbar_{m_1}^{-1}[(x_{it1}, \ldots, x_{itk} = l_1, \ldots, x_{itK})\alpha_m]\hbar_{m_2}^{-1}[(x_{it1}, \ldots, x_{itk} = l_1, \ldots, x_{itK})\beta_m] \end{aligned}$$

However, comparing to the marginal effects that are exhibited above for the Type II m2RGGMs, these effects on the expected overall healthcare cost are more complicated.

The marginal effect $\eta_k(x_{it}, \vartheta)$ in (15) can be evaluated at any particular combination of covariate values, say $X_{it} = x_{it}^{(0)}$. Therefore, the effect only represents a negligible portion of the entire population. In the most of health economics and health service studies, however, investigators would assess the marginal effect on the expected overall healthcare cost in the whole population. For this purpose, the investigators are commonly interested in the expected value of $\eta_k(x_{it}, \vartheta)$ over the population distribution of all covariates $X_{it} = (X_{it1}, X_{it2}, \ldots, X_{itK})$. This quantity, named average marginal effect, is given by

$$\bar{\eta}_k(\vartheta) = \frac{1}{T}\sum_{t=1}^{T} E_{X_{it}}[\eta_k(X_{it}, \vartheta)] = \frac{1}{T}\sum_{t=1}^{T}\int \eta_k(x_{it}, \vartheta)\mathrm{d}F_{X_{it}}(x_{it}) \tag{17}$$

where $F_{x_{it}}(\cdot)$ is the joint distribution of random vector X_{it} at the time point t, and the expectation $E_{X_{it}}[\cdot]$ is taken with respect to covariate X_{it} at the time point t. The average marginal effect (17) represents the population-average rate of marginal change (marginally with respect to the population distribution of X_{it} over time) in the expected overall healthcare cost with respect to covariate x_{itk} over time, when controlling for other factors $x_{it,-k}$. Similarly, the average elasticity and average semi-elasticity of the expected overall healthcare cost $E(y_{it}|x_{it})$ with respect to covariate x_{itk} over time can be defined as

$$\bar{e}_k(\vartheta) = \frac{1}{T}\sum_{t=1}^{T} E_{X_{it}}[e_k(X_{it}, \vartheta)]$$

and

$$\bar{s}_k(\vartheta) = \frac{1}{T}\sum_{t=1}^{T} E_{X_{it}}[s_k(X_{it}, \vartheta)]$$

respectively. When x_{itk} is categorical, the average incremental effect is given by

$$\bar{\pi}_k(\vartheta) = \frac{1}{T}\sum_{t=1}^{T} E_{X_{it}}[\pi_k(X_{it}, \vartheta)]$$

The quantity $\bar{\pi}_k(\vartheta)$ is the population-average contrast in the expected overall healthcare cost $E(y_{it}|x_{it})$ over time at the two levels of x_{itk} adjusting for all other covariates.

For the Type I and Type II m2RGGMs, marginal and incremental effects on the expected overall healthcare cost with respect to a covariate are given above. Estimators of these effects can be obtained by replacing the unknown parameters with the estimates of the parameters. When the Type II m2RGGMs are used, the marginal effects on the expected overall healthcare cost are estimated by

$$\hat{\eta}_k(x_{it}, \hat{\vartheta}) = \dot{\hbar}_{m_3}^{-1}(x_{it}'\hat{\beta}_m) \cdot \hat{\beta}_{m,k}$$

$$\hat{e}_k(x_{it}, \hat{\vartheta}) = \frac{\dot{\hbar}_{m_3}^{-1}(x'_{it}\hat{\beta}_m)}{\hbar_{m_3}^{-1}(x'_{it}\hat{\beta}_m)} x_{itk}\hat{\beta}_{m,k}$$

$$\hat{s}_k(x_{it}, \hat{\vartheta}) = \frac{\dot{\hbar}_{m_3}^{-1}(x'_{it}\hat{\beta}_m)}{\hbar_{m_3}^{-1}(x'_{it}\hat{\beta}_m)} \hat{\beta}_{m,k}$$

and

$$\hat{\pi}_k(x_{it,-k}, \hat{\vartheta}) = \hbar_{m_3}^{-1}[(x_{it1}, \ldots, x_{itk} = l_2, \ldots, x_{itK})\hat{\beta}_m] - \hbar_{m_3}^{-1}[(x_{it1}, \ldots, x_{itk} = l_1, \ldots, x_{itK})\hat{\beta}_m]$$

Here, the hat on the marginal effects indicates that they have been estimated by replacing the unknown parameters with the estimates. When the Type I m2RGGMs are used, the marginal effects on the expected overall healthcare cost are estimated by

$$\hat{\eta}_k(x_{it}, \hat{\vartheta}) = \dot{\hbar}_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)\hbar_{m_2}^{-1}(x'_{it}\hat{\beta}_m) \cdot \hat{\alpha}_{m,k} + \hbar_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)\dot{\hbar}_{m_2}^{-1}(x'_{it}\hat{\beta}_m) \cdot \hat{\beta}_{m,k}$$

$$\hat{e}_k(x_{it}, \hat{\vartheta}) = \frac{\dot{\hbar}_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)}{\hbar_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)} x_{itk}\hat{\alpha}_{m,k} + \frac{\dot{\hbar}_{m_2}^{-1}(x'_{it}\hat{\beta}_m)}{\hbar_{m_2}^{-1}(x'_{it}\hat{\beta}_m)} x_{itk}\hat{\beta}_{m,k}$$

$$\hat{s}_k(x_{it}, \hat{\vartheta}) = \frac{\dot{\hbar}_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)}{\hbar_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)} \hat{\alpha}_{m,k} + \frac{\dot{\hbar}_{m_2}^{-1}(x'_{it}\hat{\beta}_m)}{\hbar_{m_2}^{-1}(x'_{it}\hat{\beta}_m)} \hat{\beta}_{m,k}$$

and

$$\begin{aligned}\hat{\pi}_k(x_{it,-k}, \hat{\vartheta}) = {} & \hbar_{m_1}^{-1}[(x_{it1}, \ldots, x_{itk} = l_2, \ldots, x_{itK})\hat{\alpha}_m]\hbar_{m_2}^{-1}[(x_{it1}, \ldots, x_{itk} = l_2, \ldots, x_{itK})\hat{\beta}_m] \\ & - \hbar_{m_1}^{-1}[(x_{it1}, \ldots, x_{itk} = l_1, \ldots, x_{itK})\hat{\alpha}_m]\hbar_{m_2}^{-1}[(x_{it1}, \ldots, x_{itk} = l_1, \ldots, x_{itK})\hat{\beta}_m]\end{aligned}$$

Estimators of the average marginal effects can be obtained by averaging the individual marginal effects. When the Type II m2RGGMs are used, the average marginal effects on the expected overall healthcare cost are estimated by

$$\hat{\bar{\eta}}_k(\vartheta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\dot{\hbar}_{m_3}^{-1}(x'_{it}\hat{\beta}_m) \cdot \hat{\beta}_{m,k}\right\}$$

$$\hat{\bar{e}}_k(\hat{\vartheta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{\dot{\hbar}_{m_3}^{-1}(x'_{it}\hat{\beta}_m)}{\hbar_{m_3}^{-1}(x'_{it}\hat{\beta}_m)} x_{itk}\hat{\beta}_{m,k}\right\}$$

$$\hat{\bar{s}}_k(\hat{\vartheta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{\dot{\hbar}_{m_3}^{-1}(x'_{it}\hat{\beta}_m)}{\hbar_{m_3}^{-1}(x'_{it}\hat{\beta}_m)} \hat{\beta}_{m,k}\right\}$$

and

$$\bar{\pi}_k(\hat{\vartheta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\hbar_{m_3}^{-1}[(x_{it1}, \ldots, x_{itk} = l_2, \ldots, x_{itK})\hat{\beta}_m] - \hbar_{m_3}^{-1}[(x_{it1}, \ldots, x_{itk} = l_1, \ldots, x_{itK})\hat{\beta}_m]\right\}$$

Likewise, when the Type I m2RGGMs are used, the average marginal effects on the expected overall healthcare cost are estimated by

$$\hat{\bar{\eta}}_k(\vartheta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\dot{\hbar}_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)\hbar_{m_2}^{-1}(x'_{it}\hat{\beta}_m) \cdot \hat{\alpha}_{m,k} + \hbar_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)\dot{\hbar}_{m_2}^{-1}(x'_{it}\hat{\beta}_m) \cdot \hat{\beta}_{m,k}\right\}$$

$$\hat{\bar{e}}_k(\hat{\vartheta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{\dot{\hbar}_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)}{\hbar_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)} x_{itk}\hat{\alpha}_{m,k} + \frac{\dot{\hbar}_{m_2}^{-1}(x'_{it}\hat{\beta}_m)}{\hbar_{m_2}^{-1}(x'_{it}\hat{\beta}_m)} x_{itk}\hat{\beta}_{m,k}\right\}$$

$$\hat{\bar{s}}_k(\hat{\vartheta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{\dot{\hbar}_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)}{\hbar_{m_1}^{-1}(x'_{it}\hat{\alpha}_m)} \hat{\alpha}_{m,k} + \frac{\dot{\hbar}_{m_2}^{-1}(x'_{it}\hat{\beta}_m)}{\hbar_{m_2}^{-1}(x'_{it}\hat{\beta}_m)} \hat{\beta}_{m,k}\right\}$$

and

$$\hat{\bar{\pi}}_k(\hat{\vartheta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\{\hbar_{m_1}^{-1}[(x_{it1},\ldots,x_{itk}=l_2,\ldots,x_{itK})\hat{\alpha}_m] \times \hbar_{m_2}^{-1}[(x_{it1},\ldots,x_{itk}=l_2,\ldots,x_{itK})\hat{\beta}_m]$$
$$- \hbar_{m_1}^{-1}[(x_{it1},\ldots,x_{itk}=l_1,\ldots,x_{itK})\hat{\alpha}_m] \times \hbar_{m_2}^{-1}[(x_{it1},\ldots,x_{itk}=l_1,\ldots,x_{itK})\hat{\beta}_m]\}$$

In some panel studies on healthcare costs, investigators may also be interested in the marginal effects on the positive healthcare costs, besides the effects on the overall costs that includes both zero and non-zero costs. This happens when they aim to study the impact of some covariates on the healthcare expenditures given the fact that the healthcare expenditures actually occur. Marginal and incremental effects on the expected positive healthcare costs are defined in Web Appendix 1.

Variance estimation of the marginal effects and average marginal effects introduced above is essential for statistical inference procedures, such as confidence interval estimation and hypothesis testing, when using the m2RGGMs. Variance estimators for these effects, which are obtained using either the delta methods or Taylor series approximations, are derived in Web Appendix 2.

4 Maximum likelihood estimation and simulation studies

Maximum likelihood estimation of the marginalized two-part random-effects models is as demanding as the conditional two-part random-effects models. For both Type I and Type II m2RGGMs introduced in Section 2, let $\vartheta = (\alpha_m, \theta_A, \beta_m, \theta_B, \delta_m, \kappa, \theta_{AB})'$ denote the parameter vector that contains all unknown parameters. Then, the likelihood function of the models is

$$L(\vartheta) = \prod_{i=1}^{N}\int_{a_i}\int_{b_i}\left\{\prod_{t=1}^{T} P(y_{it}>0|x_{ij},a_{it})^{I(y_{it}>0)} P(y_{it}=0|x_{ij},a_{it})^{1-I(y_{it}>0)} f(y_{it}|y_{it}>0,x_{ij},b_{it})^{I(y_{it}>0)}\right\} f_{a_ib_i}(a_i,b_i)\mathrm{d}a_i\mathrm{d}b_i \tag{18}$$

in which $P(y_{it}>0|x_{ij},a_{it}) = \hbar_{c_1}^{-1}(\Delta_{it}+a_{it})$, $P(y_{it}=0|x_{ij},a_{it}) = 1-\hbar_{c_1}^{-1}(\Delta_{it}+a_{it})$, $f(y_{it}|y_{it}>0,x_{ij},b_{it}) = f_{GG}(y_{it};\kappa,\mu_{it}=\Lambda_{it}+b_{it},\sigma_{it}=\exp(X'_{it}\delta_m))$ as defined in (10), and $f_{a_ib_i}(a_i,b_i)$ is the joint distribution of $a_i=(a_{it})'$ and $b_i=(b_{it})'$. The Δ_{it} can be obtained by solving (8) and the Λ_{it} can be found in either (11) and (14). Partial derivatives of Δ_{it} and Λ_{it} with respect to the unknown parameters are given in Web Appendix 3. Adaptive Gaussian quadrature can be employed to numerically evaluate the integrals in (18). Combined with a optimization algorithm such as the Newton-Raphson or quasi-Newton, the maximum likelihood estimates of the unknown parameters can be obtained. Variance–covariance matrix of the maximum likelihood estimator $\hat{\vartheta}$ is the inverse of the Fisher information matrix, and standard errors (SEs) of the maximum likelihood estimators of the unknown parameters can be estimated by the inverse of the observed information matrix.

Simulation studies were conducted to evaluate the performance of maximum likelihood estimation for unknown parameters and marginal effects in the Type I and Type II m2RGGMs. In the first simulation study, simulation data sets were generated from a Type I m2RGGM with the form of

$$\begin{aligned}
\text{Part (I)}: \quad & \Phi^{-1}[P(Y_{it}>0|X_{it})] = \alpha_{m0} + X_{it1}\alpha_{m1}, \\
& \Phi^{-1}[P(Y_{it}>0|X_{it},a_{it})] = \Delta_{it}(X'_{it}\alpha_m, \mathrm{var}(a_{it})) + a_{i0}; \\
\text{Part (II)}: \quad & \log[E(Y_{it}|Y_{it}>0,X_{it})] = \beta_{m0} + X_{it1}\beta_{m1} + X_{it2}\beta_{m2} + X_{it3}\beta_{m3}, \\
& (\mu_{it}|Y_{it}>0,X_{it},b_{it}) = \Lambda_{it}(X'_{it}\beta_m, X'_{it}\delta_m, \mathrm{var}(b_{it})) + b_{i0}, \\
& \log(\sigma_{it}|Y_{it}>0,X_{it}) = \delta_{m0} + X_{it1}\delta_{m1}
\end{aligned} \tag{19}$$

and from a Type II m2RGGM with the form of

$$\begin{aligned}
\text{Part (I)}: \quad & \Phi^{-1}[P(Y_{it}>0|X_{it})] = \alpha_{m0} + X_{it1}\alpha_{m1}, \\
& \Phi^{-1}[P(Y_{it}>0|X_{it},a_{it})] = \Delta_{it}(X'_{it}\alpha_m, \mathrm{var}(a_{it})) + a_{i0};
\end{aligned}$$

$$\begin{aligned}\text{Part (II)}: \quad & \log[E(Y_{it}|X_{it})] = \beta_{m0} + X_{it1}\beta_{m1} + X_{it2}\beta_{m2} + X_{it3}\beta_{m3},\\ & (\mu_{it}|Y_{it} > 0, X_{it}, b_{it}) = \Lambda_{it}(X'_{it}\alpha_m, X'_{it}\beta_m, X'_{it}\delta_m, \mathrm{var}(b_{it})) + b_{i0},\\ & \log(\sigma_{it}|Y_{it} > 0, X_{it}) = \delta_{m0} + X_{it1}\delta_{m1} \end{aligned} \tag{20}$$

in which the time-invariant covariate X_{it1} followed a Bernoulli distribution with $P(X_{it1} = 1) = P(X_{it1} = 0) = 0.5$, the time-invariant covariate X_{it2} followed a uniform distribution Uniform(0, 1), $X_{it3} = t/4$ with $t = 1, 2, 3, 4$, $\alpha_m = (\alpha_{m0}, \alpha_{m1})' = (-0.5, -0.5)'$, $\beta_m = (\beta_{m0}, \beta_{m1}, \beta_{m2}, \beta_{m3})' = (2, 1, 1, 1)'$, $\delta_m = (\delta_{m0}, \delta_{m1})' = (-0.2, -0.4)'$, and the correlated random intercepts a_{i0} and b_{i0} followed a bivariate normal distribution with $\mathrm{var}(a_{i0}) = \mathrm{var}(b_{i0}) = 0.5$ and $\mathrm{corr}(a_{i0}, b_{i0}) = 0.5$. The shape parameter κ was set to be 2 in both (19) and (20). This simulation study was conducted with 100 simulation data sets, each of which consists of $N = 2000$ individuals ($j = 1, 2, \ldots, N$). For the data sets simulated from (19) and (20), 76.659% of the simulated healthcare expenditures were zero. In (19), the average positive healthcare expenditure was 50.811 with a standard deviation of 84.396. In (20), the average overall healthcare expenditure was 60.261 with a standard deviation of 269.594. The maximum likelihood estimates of unknown parameters were obtained by using the SAS procedure NLMIXED for numerical evaluation and optimization of likelihood (18). Table 1 reports the true values of parameters, the means of parameter estimates, the standard deviations, the means of SEs, and the coverage probabilities of the corresponding 95% confidence interval for the Type I and Type II m2RGGMs. The simulation results in Table 1 demonstrate that the empirical biases of the parameter estimates are negligible. The standard deviations of the parameter estimates are small enough to conclude that the estimates are unbiased. The average SEs are generally close to standard deviations. The coverage probabilities for most of the parameters are acceptable comparing to the nominal level 0.95. But, for the heteroscedastic variance parameters δ_{m0}, δ_{m1} and κ, the coverage probabilities are lower than 0.95, which may be caused by the small finite sample bias in both parameter estimates and SEs. We investigated the performance of estimating the incremental effect of X_{it1} and the marginal effects (including elasticity and semi-elasticity) of X_{it2} on the expected overall healthcare expenditure at $X_{it2} = 0.5$ and $t = 4$ for the Type II m2RGGM (20), as well as the incremental effect of X_{it1} and the marginal effects of X_{it2} on the expected positive healthcare expenditure at $X_{it2} = 0.5$ and $t = 4$ for the Type I m2RGGM (19). Estimates of the average incremental effect of X_{it1} and the average marginal effects of X_{it2} on the expected overall healthcare expenditure for (20) and on the expected positive healthcare expenditure for (19) were also evaluated. Table 2 reports the true values of (average) marginal and incremental effects, the means of the effect estimates, the standard deviations, the means of SEs, and the coverage probabilities of the corresponding 95% confidence interval for the Type I and Type II m2RGGMs. Table 2 shows that the estimates of the marginal and incremental effects in (19) and (20) perform well as the parameter estimates do, and the empirical biases of the effect estimates are negligible as well.

Two additional simulation studies were conducted to evaluate the performance of maximum likelihood estimation in the Type I and Type II m2RGGMs with four random effects and to evaluate the impact of

Table 1. True values, parameter estimates, standard deviations (SD), standard errors (SE), and coverage probabilities (CP) for the Type I and Type II m2RGGMs in the simulation study.

	Type I m2RGGM					Type II m2RGGM				
Estimand	True value	Estimate	SD	SE	CP	True value	Estimate	SD	SE	CP
α_{m0}	−0.5	−0.501	0.026	0.026	0.94	−0.5	−0.500	0.024	0.026	0.99
α_{m1}	−0.5	−0.500	0.040	0.039	0.93	−0.5	−0.496	0.036	0.039	0.96
β_{m0}	2	2.001	0.077	0.091	0.98	2	1.984	0.100	0.098	0.93
β_{m1}	1	1.003	0.064	0.065	0.96	1	1.003	0.080	0.091	0.98
β_{m2}	1	0.995	0.116	0.107	0.93	1	1.007	0.111	0.107	0.94
β_{m3}	1	0.999	0.081	0.087	0.95	1	1.021	0.094	0.087	0.94
δ_{m0}	−0.2	−0.212	0.185	0.099	0.82	−0.2	−0.264	0.223	0.098	0.81
δ_{m1}	−0.4	−0.398	0.068	0.058	0.88	−0.4	−0.395	0.061	0.058	0.92
$\mathrm{var}(a_{i0})$	0.5	0.502	0.044	0.047	0.97	0.5	0.500	0.048	0.046	0.95
$\mathrm{var}(b_{i0})$	0.5	0.486	0.063	0.059	0.92	0.5	0.482	0.059	0.056	0.92
$\mathrm{corr}(a_{i0}, b_{i0})$	0.5	0.506	0.088	0.078	0.90	0.5	0.486	0.101	0.078	0.87
κ	2	2.062	0.531	0.248	0.79	2	2.212	0.664	0.254	0.80

Table 2. True values, estimates, standard deviations (SD), standard errors (SE), and coverage probabilities (CP) of (average) incremental and marginal effects for the Type I and Type II m2RGGMs in the simulation study.

	Type I m2RGGM						Type II m2RGGM					
x_{it1}	Estimand	True value	Estimate	SD	SE	CP	Estimand	True value	Estimate	SD	SE	CP
	$\pi_1(x_{it},\vartheta)$	56.902	57.128	4.902	5.557	0.98	$\pi_1(x_{it},\vartheta)$	56.902	57.955	7.568	7.446	0.95
$x_1=0$	$\eta_2(x_{it},\vartheta)$	33.115	32.958	4.565	4.069	0.89	$\eta_2(x_{it},\vartheta)$	33.115	33.779	4.920	4.280	0.91
$x_1=1$	$\eta_2(x_{it},\vartheta)$	90.017	89.731	11.541	11.564	0.95	$\eta_2(x_{it},\vartheta)$	90.017	92.268	15.018	12.582	0.91
$x_1=0$	$e_2(x_{it},\vartheta)$	0.5	0.498	0.058	0.054	0.93	$e_2(x_{it},\vartheta)$	0.5	0.503	0.056	0.053	0.94
$x_1=1$	$e_2(x_{it},\vartheta)$	0.5	0.498	0.058	0.054	0.93	$e_2(x_{it},\vartheta)$	0.5	0.503	0.056	0.053	0.94
$x_1=0$	$s_2(x_{it},\vartheta)$	1	0.995	0.116	0.107	0.93	$s_2(x_{it},\vartheta)$	1	1.007	0.111	0.107	0.94
$x_1=1$	$s_2(x_{it},\vartheta)$	1	0.995	0.116	0.107	0.93	$s_2(x_{it},\vartheta)$	1	1.007	0.111	0.107	0.94
	$\bar{\pi}_k(\vartheta)$	42.367	42.603	3.801	4.111	0.96	$\bar{\pi}_k(\vartheta)$	42.367	42.912	5.347	5.366	0.96
	$\bar{\eta}_k(\vartheta)$	39.044	39.008	5.475	5.073	0.95	$\bar{\eta}_k(\vartheta)$	45.840	46.709	7.096	6.158	0.93
	$\bar{e}_k(\vartheta)$	0.5	0.498	0.058	0.054	0.95	$\bar{e}_k(\vartheta)$	0.5	0.504	0.056	0.054	0.94
	$\bar{s}_k(\vartheta)$	1	0.995	0.116	0.107	0.93	$\bar{s}_k(\vartheta)$	1	1.007	0.111	0.107	0.94

random effects misspecification on maximum likelihood estimation. The results of these simulation studies are reported in Web Appendix 4 and Web Appendix 5, respectively.

5 Analysis of healthcare cost data from China

In this section, a set of healthcare cost panel data is taken from the CHNS[19,13] and analyzed using the marginalized and conditional two-part random-effects models introduced in Section 2.

5.1 Data description

As a collaborative project between the Carolina Population Center at the University of North Carolina at Chapel Hill and the National Institute for Nutrition and Health at the Chinese Center for Disease Control and Prevention, the CHNS is designed to investigate the effects of health and nutrition programs implemented by Chinese governments throughout the past one and half decades. The primary goal of the CHNS is to determine how the social and economic transformation in China impacts the health and nutritional status of the population during China's rapid economic ascendance. The first wave of the CHNS was conducted in 1989, followed by eight subsequent waves in 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011. During each wave, the survey implemented a multistage, random cluster process of data collection to draw a sample of households and the individuals within. The household and individual surveys contain several modules on respondent demographics, health, nutrition, and income. Until 2011, the survey covers a total number of nine provinces that vary substantially in geographical factors, economic development, public resources, and health indicators. The names of the nine provinces are Liaoning, Heilongjiang, Guangxi, Guizhou, Henan, Hubei, Hunan, Jiangsu, and Shandong. CHNS has been expanded to include a health services section, which contains detailed data on insurance coverage, medical providers, healthcare utilization, and healthcare costs in the past four weeks. Questions about accessibility to healthcare facilities, time and travel costs to receive medical care, and perceived quality of medical care are also asked. Information on illnesses and on all uses of the health system during the previous month is collected for all household members.

The primary dependent variable in our analysis regarding the healthcare expenditures in the CHNS data is the total medical expenditure of each individual during the previous four weeks of the survey (denoted by Y_{it} for individual i at the tth survey year). The key explanatory variables include whether the individual came from a rural area (this variable, although denoted by X_{it1}, is not time-varying with 1 representing the individual came from a rural area and 0 representing urban area), baseline age of the individual (denoted by X_{it2} but not time-varying), whether the individual had at least one type of diseases listed in the survey questionnaire (denoted by X_{it3}; 1 represents the individual had either diabetes, hypertension, myocardial infarction, apoplexy, bone fracture or asthma and 0 represents otherwise), nature logarithm of one plus individual income (denoted by X_{it4}), health insurance status (denoted by X_{it5}; 1 represents the individual was covered by at least one type of health insurance and 0 represents otherwise), and survey wave (i.e., survey year; denoted by X_{it6}). The total medical expenditure and individual income, both in Chinese yuan (yuan is the official currency of China), were converted to the 2011 price level according to the annual customer price indices published by the National Bureau of Statistics of China.[20] Eight dummy variables were created to represent eight provinces (U_{i1} for Heilongjiang U_{i2} for Jiangsu, U_{i3} for Shandong, U_{i4} for Henan, U_{i5} for Hubei, U_{i6} for Hunan, U_{i7} for Guangxi, U_{i8} for Guizhou), with Liaoning as the reference province. At the end of 1998, Chinese government enacted an mandatory insurance program, the UEBMI, to reform the urban employee medical insurance system.[21] In 2003, China also launched a new health insurance system, the New Cooperative Medical Scheme (NCMS), in rural areas where the majority of the population does not have any kind of health insurance before the NCMS reform.[22] In addition, the CHNS has adopted a set of unified survey questionnaires since the year of 2004. Considering these two important reforms that may dramatically affect medical expenses and considering the change of survey questionnaires in the CHNS, we include in our analysis 2639 adults who participated the CHNS after the enactment of UEBMI and NCMS in four waves from 2004 to 2011.

5.2 Statistical modeling

Our modeling strategy in analyzing the medical expenditure panel data collected from the CHNS is to take the advantages of the proposed Type I and Type II m2RGGMs and provide instant marginal inference by estimating the marginal effects of the key independent variables with respect to the total medical expenditures. The Type I m2RGGM that was used to analyze the CHNS data takes the form of

$$\begin{aligned}
\text{Part (I)}: \quad & \Phi^{-1}[P(Y_{it} > 0 | X_{it})] = \alpha_{m0} + \sum_{k=1}^{6} X_{itk}\alpha_{mk} + \sum_{k=1}^{8} U_{ik}\alpha_{m(k+6)}, \\
& \Phi^{-1}[P(Y_{it} > 0 | X_{it}, a_{it})] = \Delta_{it}(X'_{it}\alpha_m, \text{var}(a_{it})) + a_{i0}; \\
\text{Part (II)}: \quad & \log[E(Y_{it} | Y_{it} > 0, X_{it})] = \beta_{m0} + \sum_{k=1}^{6} X_{itk}\beta_{mk} + \sum_{k=1}^{8} U_{ik}\beta_{m(k+6)}, \\
& (\mu_{it} | Y_{it} > 0, X_{it}, b_{it}) = \Lambda_{it}(X'_{it}\beta_m, X'_{it}\delta_m, \text{var}(b_{it})) + b_{i0}, \\
& \log(\sigma_{it} | Y_{it} > 0, X_{it}) = \delta_{m0} + \sum_{k=1}^{6} X_{itk}\delta_{mk} + \sum_{k=1}^{8} U_{ik}\delta_{m(k+6)}
\end{aligned} \tag{21}$$

The Type II m2RGGM that was used to analyze the CHNS data takes the form of

$$\begin{aligned}
\text{Part (I)}: \quad & \Phi^{-1}[P(Y_{it} > 0 | X_{it})] = \alpha_{m0} + \sum_{k=1}^{6} X_{itk}\alpha_{mk} + \sum_{k=1}^{8} U_{ik}\alpha_{m(k+6)}, \\
& \Phi^{-1}[P(Y_{it} > 0 | X_{it}, a_{it})] = \Delta_{it}(X'_{it}\alpha_m, \text{var}(a_{it})) + a_{i0}; \\
\text{Part (II)}: \quad & \log[E(Y_{it} | X_{it})] = \beta_{m0} + \sum_{k=1}^{6} X_{itk}\beta_{mk} + \sum_{k=1}^{8} U_{ik}\beta_{m(k+6)}, \\
& (\mu_{it} | Y_{it} > 0, X_{it}, b_{it}) = \Lambda_{it}(X'_{it}\alpha_m, X'_{it}\beta_m, X'_{it}\delta_m, \text{var}(b_{it})) + b_{i0}, \\
& \log(\sigma_{it} | Y_{it} > 0, X_{it}) = \delta_{m0} + \sum_{k=1}^{6} X_{itk}\delta_{mk} + \sum_{k=1}^{8} U_{ik}\delta_{m(k+6)}
\end{aligned} \tag{22}$$

Type I m2RGGM (21) and Type II m2RGGM (22) are specific forms of the proposed Type I and Type II m2RGGMs in Section 2, where $\hbar_{m_1}(\cdot) = \hbar_{c_1}(\cdot) = \Phi^{-1}(\cdot)$, $\hbar_{m_2}(\cdot) = \hbar_{m_3}(\cdot) = \log(\cdot)$. In addition, $\alpha_m = (\alpha_{m0}, \alpha_{m1}, \ldots, \alpha_{m,14})'$, $\beta_m = (\beta_{m0}, \beta_{m1}, \ldots, \beta_{m,14})'$, and $\delta_m = (\delta_{m0}, \delta_{m1}, \ldots, \delta_{m,14})'$ contain the regression coefficients for the intercepts and key independent variables, and a_{i0} and b_{i0} are correlated random intercepts that follow a bivariate normal distribution $N_2(0, \Sigma)$ with $\text{var}(a_{i0}) = \theta_A^2$, $\text{var}(b_{i0}) = \theta_B^2$, and $\text{corr}(a_{i0}, b_{i0}) = \theta_{AB}$. The maximum likelihood estimates of the unknown parameters in (21) and (22) were obtained by maximizing (18), for which the SAS procedure NLMIXED was used to perform numerical evaluation and optimization. The SAS program for the maximum likelihood estimation is demonstrated in Web Appendix 6. To supply a reliable initial value to the numerical optimization in NLMIXED, we implemented a two-step procedure. For both Type I m2RGGM (21) and Type II m2RGGM (22), a marginal likelihood function for Part (I)

$$L_1(\alpha_m, \theta_A) = \prod_{i=1}^{N} \int_{a_{i0}} \left\{ \prod_{t=1}^{T} [\Phi(\Delta_{it} + a_{i0})]^{I(y_{it} > 0)} [1 - \Phi(\Delta_{it} + a_{i0})]^{1 - I(y_{it} > 0)} \right\} f_{a_{i0}}(a_{i0}) \mathrm{d}a_{i0} \tag{23}$$

was first maximized to provide an initial value for (α_m, θ_A), denoted by $(\tilde{\alpha}_m, \tilde{\theta}_A)$. Then, $(\tilde{\alpha}_m, \tilde{\theta}_A)$ was inserted into $L(\vartheta)$ in (18) and $(\tilde{\beta}_m, \tilde{\theta}_B, \tilde{\delta}, \tilde{\kappa}, \tilde{\theta}_{AB}) = \arg\max_{(\beta_m, \theta_B, \delta, \kappa, \theta_{AB})} L(\tilde{\alpha}_m, \tilde{\theta}_A, \beta_m, \theta_B, \delta, \kappa, \theta_{AB})$ was solved. The estimates $(\tilde{\beta}_m, \tilde{\theta}_B, \tilde{\delta}, \tilde{\kappa}, \tilde{\theta}_{AB})$ were classified as maximum pseudo-likelihood estimates by Gong and Samaniego.[23] After these two steps, we take $\tilde{\vartheta} = (\tilde{\alpha}_m, \tilde{\theta}_A, \tilde{\beta}_m, \tilde{\theta}_B, \tilde{\delta}, \tilde{\kappa}, \tilde{\theta}_{AB})$ as the initial value. Here, we acknowledge that a major limitation of the proposed Type I m2RGGMs and Type II m2RGGMs in Section 2 is that the computational burden of maximizing the likelihood will sharply increase as the number of random coefficients goes up. Please see Section 6 for more discussion and potential solutions. For the purpose of comparison, a c2RGGM with a form of

$$\begin{aligned} &\text{Part (I)}: \quad \Phi^{-1}[P(Y_{it} > 0 | X_{it}, a_{it})] = \alpha_{c0} + \sum_{k=1}^{6} X_{itk}\alpha_{ck} + \sum_{k=1}^{8} U_{ik}\alpha_{c(k+6)} + a_{i0}; \\ &\text{Part (II)}: \quad (\mu_{it} | Y_{it} > 0, X_{it}, b_{it}) = \beta_{c0} + \sum_{k=1}^{6} X_{itk}\beta_{ck} + \sum_{k=1}^{8} U_{ik}\beta_{c(k+6)}, + b_{i0}, \\ &\qquad \log(\sigma_{it} | Y_{it} > 0, X_{it}) = \delta_{c0} + \sum_{k=1}^{6} X_{itk}\delta_{ck} + \sum_{k=1}^{8} U_{ik}\delta_{c(k+6)} \end{aligned} \tag{24}$$

was taken to analyze the CHNS data as well.

5.3 Empirical results

The analysis results from fitting the Type I m2RGGM to the CHNS data are summarized in Table 3. The table reports the estimates, estimated SEs and p values associated with the regression parameters in Parts (I) and (II) of the full model that includes all covariates and the reduced model in which the covariates with an insignificant p value are excluded. Part (II) in Table 3 demonstrates that rural status, age, disease status, log income (logarithm of one plus income), and survey year have statistically significant impact on the expected positive healthcare cost. Insurance status is not a statistically significant factor on the expected positive healthcare cost, which implies that whether or not the individual is insured does not affect the amount of the healthcare expenditure while controlling for other factors. In Part (II) of Table 3, two provinces, Shandong and Guizhou, exhibit a significant departure from the baseline province, Liaoning, in the expected amount of positive healthcare cost. Part (I) in Table 3 demonstrates that rural status, age, disease status, and log income have statistically significant impact on the healthcare utilization (i.e., the probability of observing a positive healthcare cost or the presence of the healthcare expenditure). The negative estimated coefficient for rural status in Part (I) implies the individuals from rural areas accessed the healthcare system less frequently than the ones from urban areas. This is consistent with the urban–rural inequality in healthcare system in China as China is still on its way of urbanization. The positive estimated coefficients for age and disease status show the increase of healthcare utilization among the elders and the diseased. It is interesting to observe the healthcare utilization was decreased as the log income increased. Insurance status is still not a statistically significant factor on the healthcare utilization. The insignificance in survey year indicates that the chance of utilizing the healthcare system for an individual did not vary over the years in China while controlling for other factors. In Part (I) of Table 3, five provinces, Heilongjiang, Shandong, Henan, Hubei, and Guizhou, exhibit a significant departure from the baseline province in the healthcare utilization, which is an evidence of heterogeneity in the use of health care system across different provinces.

Heteroscedasticity among the positive healthcare costs is confirmed by the significance of disease status and the province of Shandong (comparing to the baseline province) in the heteroscedastic variance model of the Type I m2RGGM. There is a strong evidence that the shape parameter κ ($p < 0.0001$) is not equal to zero, showing that the generalized Gamma model in Part (II) is superior to the traditional lognormal model. The estimated variance components $\hat{\theta}_A$ and $\hat{\theta}_B$ in Table 3 with highly significant p values ($p < 0.0001$) suggest that unexplained

Table 3. Parameter estimates, estimated standard errors, and *p* values from fitting the Type I m2RGGM to the CHNS data.

		Full model			Reduced model		
	Parameter	Estimate	SE	*p* value	Estimate	SE	*p* value
Part (I)	Intercept (α_{m0})	−1.547	0.108	<.0001	−1.583	0.097	<.0001
	Rural (α_{m1})	−0.144	0.040	0.000	−0.135	0.040	0.001
	Age (α_{m2})	0.010	0.001	<.0001	0.010	0.001	<.0001
	Disease (α_{m3})	0.560	0.042	<.0001	0.561	0.041	<.0001
	Insurance (α_{m4})	0.075	0.044	0.091			
	log(1+Income) (α_{m5})	−0.018	0.004	<.0001	−0.016	0.004	0.000
	Wave (α_{m6})	−0.010	0.007	0.180			
	Heilongjiang (α_{m7})	−0.500	0.078	<.0001	−0.453	0.067	<.0001
	Jiangsu (α_{m8})	0.001	0.069	0.986			
	Shandong (α_{m9})	−0.356	0.078	<.0001	−0.298	0.067	<.0001
	Henan (α_{m10})	−0.234	0.073	0.001	−0.185	0.060	0.002
	Hubei (α_{m11})	−0.184	0.071	0.009	−0.148	0.058	0.011
	Hunan (α_{m12})	−0.121	0.074	0.101			
	Guangxi (α_{m13})	−0.050	0.074	0.501			
	Guizhou (α_{m14})	−0.215	0.071	0.002	−0.173	0.057	0.003
Part (II)	Intercept (β_{m0})	4.939	0.385	<.0001	4.914	0.343	<.0001
	Rural (β_{m1})	−0.284	0.142	0.045	−0.323	0.137	0.018
	Age (β_{m2})	0.010	0.005	0.035	0.013	0.005	0.008
	Disease (β_{m3})	0.842	0.131	<.0001	0.832	0.129	<.0001
	Insurance (β_{m4})	−0.168	0.145	0.246			
	log(1+Income) (β_{m5})	0.037	0.015	0.017	0.029	0.014	0.045
	Wave (β_{m6})	0.121	0.023	<.0001	0.104	0.017	<.0001
	Heilongjiang (β_{m7})	0.169	0.292	0.562			
	Jiangsu (β_{m8})	−0.039	0.225	0.863			
	Shandong (β_{m9})	−0.614	0.271	0.024	−0.834	0.218	0.000
	Henan (β_{m10})	−0.037	0.282	0.896			
	Hubei (β_{m11})	0.215	0.239	0.369			
	Hunan (β_{m12})	0.347	0.246	0.158			
	Guangxi (β_{m13})	−0.113	0.247	0.647			
	Guizhou (β_{m14})	−0.485	0.237	0.041	−0.573	0.189	0.003
Heteroscedasticity	Intercept (δ_{m0})	−0.523	0.224	0.020	−0.596	0.067	<.0001
	Rural (δ_{m1})	−0.124	0.087	0.156			
	Age (δ_{m2})	0.000	0.003	0.920			
	Disease (δ_{m3})	0.191	0.084	0.023	0.169	0.078	0.030
	Insurance (δ_{m4})	−0.109	0.113	0.331			
	log(1+Income) (δ_{m5})	0.012	0.010	0.216			
	Wave (δ_{m6})	0.020	0.019	0.291			
	Heilongjiang (δ_{m7})	−0.188	0.177	0.288			
	Jiangsu (δ_{m8})	−0.181	0.131	0.168			
	Shandong (δ_{m9})	−0.375	0.177	0.035	−0.348	0.145	0.017
	Henan (δ_{m10})	0.219	0.152	0.151			
	Hubei (δ_{m11})	−0.116	0.138	0.401			
	Hunan (δ_{m12})	−0.146	0.149	0.329			
	Guangxi (δ_{m13})	0.030	0.137	0.826			
	Guizhou (δ_{m14})	−0.149	0.142	0.292			
	Kappa (κ)	2.747	0.215	<.0001	2.753	0.195	<.0001
Variance components	θ_A	0.513	0.034	<.0001	0.504	0.034	<.0001
	θ_B	1.523	0.047	<.0001	1.547	0.044	<.0001
	θ_{AB}	0.243	0.090	0.007	0.219	0.078	0.005

heterogeneity may exist. The estimated correlation coefficient $\hat{\theta}_{AB}$ is significantly positive ($p < 0.01$) suggests a strong positive correlation between the healthcare cost and the healthcare utilization.

Table 4 exhibits the analysis results from fitting the Type II m2RGGM to the CHNS data. Part (II) in Table 4 demonstrates that rural status, age, disease status, and survey year have statistically significant impact on the expected overall healthcare cost. The negative estimated coefficient for rural status implies that the overall healthcare expenditure among the individuals from rural areas is significantly less than urban areas. The positive estimated coefficients in age and disease show the overall healthcare costs increase among the elders and the diseased. According to the estimated coefficient in survey year and its p value ($p < 0.0001$), the overall healthcare costs increase significantly from 2004 to 2011 in China. In Part (II) of Table 4, neither insurance status nor log income is shown to be a statistically significant factor on the expected overall healthcare cost. Two provinces, Shandong and Guizhou, exhibit a significant departure from the baseline province in the positive healthcare costs, which is consistent with the analysis in the Type I m2RGGM. As in Table 3, Part (I) in Table 4 demonstrates that rural status, age, disease status, and log income have statistically significant impact on the healthcare utilization, and insurance status and survey year are not statistically significant, and five provinces, Heilongjiang, Shandong, Henan, Hubei, and Guizhou, exhibit a significant departure from the baseline province in the healthcare utilization is confirmed by the significance of disease status and the province of Shandong in the heteroscedastic variance model of the Type II m2RGGM. There is also a strong evidence that the shape parameter κ ($p < 0.0001$) is not equal to zero in the Type II m2RGGM. As in Table 3, the estimated variance components $\hat{\theta}_A$ and $\hat{\theta}_B$ in Table 4 are highly significant p values ($p < 0.0001$) as the evidence of unexplained heterogeneity. The estimated correlation coefficient $\hat{\theta}_{AB}$ is significantly positive ($p = 0.038$). This confirms a strong positive correlation between the healthcare cost and the healthcare utilization. For the purpose of comparison, the analysis results from fitting the c2RGGM to the CHNS data are also summarized in Table 5.

The likelihood ratio test statistics that serve to compare the reduced model to the full model were calculated for two pairs in Tables 3 and 4. The corresponding p values are both less than 0.0001, suggesting the reduced models are generally preferred. Furthermore, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) values of the reduced model are less than the values of the full model in both tables. However, this is only simple model comparison and we are not committing to developing any sophisticated model selection procedure for the m2RRGGMs since it is beyond the scope of this article.

5.4 Estimation of marginal effects

For Type II m2RGGM (22), we examine the incremental effect of rural status $\pi_1(x_{it}, \vartheta)$, the marginal effect of age $\eta_2(x_{it}, \vartheta)$, and the incremental effect of disease status $\pi_3(x_{it}, \vartheta)$ on the expected overall healthcare cost. These effects are estimated under the reduced model in Table 4 in order to eliminate the interference from the insignificant covariates. Table 6 reports the estimates and the SEs of the effects at various combinations of rural and disease status. To avoid reporting a large number of redundant tables, Table 6 only reports the effect estimates for the hypothetic individuals whose age is equal to the sample average of the age variable (i.e., $x_{it2} = 60.839$) at the year of 2011 and for the baseline province, although the effects are estimated and examined for other survey years and significant provinces in the reduced model as well. It is observed from $\hat{\pi}_1(x_{it}, \hat{\vartheta})$ that, for the individuals from Liaoning province with an average age, their healthcare expenditures when living in urban areas in 2011 were 31.989 (SE: 14.181) yuan higher than living in rural areas if they are non-diseased, whereas this difference was 138.404 (60.169) yuan if they are diseased. The estimates of $\hat{\pi}_3(x_{it}, \hat{\vartheta})$ show that for such individuals the healthcare expenditures among the diseased were 235.995 (44.540) yuan higher in 2011 than the non-diseased when living in rural areas, whereas this difference was 342.411 (66.801) yuan if they lived in urban areas. The largest marginal effect of age, $\hat{\eta}_2(x_{it}, \hat{\vartheta}) = 12.282$ (2.957), occurred among the diseased living in urban areas. This quantifies the marginal change of the expected overall healthcare cost among the subpopulation when age is increased by a small amount while holding other factors. We also estimate the average incremental effect of rural status $\bar{\pi}_1(\vartheta)$, the average marginal effect of age $\bar{\eta}_2(\vartheta)$, and the average incremental effect of disease status $\bar{\pi}_3(\vartheta)$ on the expected overall healthcare cost. The estimated average marginal effect of age on the expected overall healthcare cost is 2.500 (SE: 0.599). The estimated average incremental effects of rural status and disease status are−35.461 (15.177) and 174.923 (28.804), respectively. The estimated average semi-elasticity $\bar{s}_2(\vartheta)$ of age is instantly given by the estimated coefficient of age, which is 0.028 (0.005). The interpretation of this estimated average semi-elasticity is that there is an average change of 2.8% in the expected overall healthcare cost when age is increased by one.

For Type I m2RGGM (21), we examine the incremental effect of rural status $\pi_1(x_{it}, \vartheta)$, the marginal effect of age $\eta_2(x_{it}, \vartheta)$, the incremental effect of disease status $\pi_3(x_{it}, \vartheta)$, and the marginal effect of log income $\eta_5(x_{it}, \vartheta)$ on

Table 4. Parameter estimates, estimated standard errors, and *p* values from fitting the Type II m2RGGM to the CHNS data.

		Full model			Reduced model		
	Parameter	Estimate	SE	*p* value	Estimate	SE	*p* value
Part (I)	Intercept (α_{m0})	−1.529	0.106	<.0001	−1.544	0.095	<.0001
	Rural (α_{m1})	−0.118	0.040	0.003	−0.121	0.039	0.002
	Age (α_{m2})	0.009	0.001	<.0001	0.009	0.001	<.0001
	Disease (α_{m3})	0.561	0.041	<.0001	0.570	0.041	<.0001
	Insurance (α_{m4})	0.113	0.044	0.011			
	log(Income+1) (α_{m5})	−0.018	0.004	<.0001	−0.017	0.004	<.0001
	Wave (α_{m6})	−0.015	0.007	0.041			
	Heilongjiang (α_{m7})	−0.481	0.077	<.0001	−0.437	0.061	<.0001
	Jiangsu (α_{m8})	−0.026	0.068	0.707			
	Shandong (α_{m9})	−0.294	0.076	0.000	−0.301	0.066	<.0001
	Henan (α_{m10})	−0.244	0.072	0.001	−0.202	0.055	0.000
	Hubei (α_{m11})	−0.207	0.070	0.003	−0.181	0.053	0.001
	Hunan (α_{m12})	−0.133	0.073	0.069			
	Guangxi (α_{m13})	−0.044	0.073	0.541			
	Guizhou (α_{m14})	−0.210	0.069	0.002	−0.212	0.057	0.000
Part (II)	Intercept (β_{m0})	2.070	0.443	<.0001	2.085	0.378	<.0001
	Rural (β_{m1})	−0.345	0.163	0.035	−0.372	0.145	0.011
	Age (β_{m2})	0.026	0.006	<.0001	0.028	0.005	<.0001
	Disease (β_{m3})	1.473	0.148	<.0001	1.465	0.137	<.0001
	Insurance (β_{m4})	0.040	0.170	0.815			
	log(Income+1) (β_{m5})	0.019	0.018	0.279			
	Wave (β_{m6})	0.101	0.027	0.000	0.109	0.018	<.0001
	Heilongjiang (β_{m7})	−0.633	0.331	0.056			
	Jiangsu (β_{m8})	−0.186	0.257	0.469			
	Shandong (β_{m9})	−1.122	0.306	0.000	−1.105	0.254	<.0001
	Henan (β_{m10})	−0.353	0.309	0.253			
	Hubei (β_{m11})	−0.118	0.277	0.670			
	Hunan (β_{m12})	0.111	0.285	0.696			
	Guangxi (β_{m13})	−0.159	0.291	0.585			
	Guizhou (β_{m14})	−0.812	0.279	0.004	−0.808	0.220	0.000
Heteroscedasticity	Intercept (δ_{m0})	−0.397	0.236	0.092	−0.570	0.055	<.0001
	Rural (δ_{m1})	−0.047	0.091	0.601			
	Age (δ_{m2})	−0.003	0.003	0.347			
	Disease (δ_{m3})	0.060	0.090	0.502			
	Insurance (δ_{m4})	−0.078	0.121	0.517			
	log(Income+1) (δ_{m5})	0.016	0.010	0.102			
	Wave (δ_{m6})	0.020	0.020	0.319			
	Heilongjiang (δ_{m7})	−0.196	0.182	0.280			
	Jiangsu (δ_{m8})	−0.212	0.136	0.120			
	Shandong (δ_{m9})	−0.502	0.176	0.004	−0.483	0.172	0.005
	Henan (δ_{m10})	0.242	0.149	0.106			
	Hubei (δ_{m11})	−0.097	0.143	0.495			
	Hunan (δ_{m12})	−0.031	0.152	0.838			
	Guangxi (δ_{m13})	0.021	0.145	0.884			
	Guizhou (δ_{m14})	−0.180	0.150	0.232			
	Kappa (κ)	2.929	0.233	<.0001	2.903	0.185	<.0001
Variance components	θ_A	0.477	0.034	<.0001	0.497	0.034	<.0001
	θ_B	1.526	0.055	<.0001	1.562	0.052	<.0001
	θ_{AB}	0.318	0.139	0.023	0.133	0.064	0.038

Table 5. Parameter estimates, estimated standard errors, and *p* values from fitting the c2RGGM to the CHNS data.

		Full model			Reduced model		
	Parameter	Estimate	SE	*p* value	Estimate	SE	*p* value
Part (I)	Intercept (α_{c0})	−1.847	0.125	<.0001	−1.861	0.112	<.0001
	Rural (α_{c1})	−0.117	0.045	0.010	−0.137	0.045	0.002
	Age (α_{c2})	0.013	0.002	<.0001	0.012	0.002	<.0001
	Disease (α_{c3})	0.607	0.047	<.0001	0.601	0.046	<.0001
	Insurance (α_{c4})	0.086	0.050	0.084			
	log(1+Income) (α_{c5})	−0.019	0.005	<.0001	−0.017	0.005	<.0001
	Wave (α_{c6})	−0.010	0.008	0.234			
	Heilongjiang (α_{c7})	−0.603	0.089	<.0001	−0.525	0.076	<.0001
	Jiangsu (α_{c8})	−0.011	0.077	0.883			
	Shandong (α_{c9})	−0.394	0.088	<.0001	−0.338	0.075	<.0001
	Henan (α_{c10})	−0.314	0.082	<.0001	−0.223	0.068	0.001
	Hubei (α_{c11})	−0.235	0.080	0.003	−0.179	0.066	0.006
	Hunan (α_{c12})	−0.141	0.083	0.089			
	Guangxi (α_{c13})	−0.051	0.083	0.537			
	Guizhou (α_{c14})	−0.252	0.079	0.001	−0.214	0.064	0.001
Part (II)	Intercept (β_{c0})	3.443	0.418	<.0001	3.829	0.347	<.0001
	Rural (β_{c1})	−0.221	0.147	0.132			
	Age (β_{c2})	0.021	0.005	<.0001	0.015	0.005	0.003
	Disease (β_{c3})	0.813	0.139	<.0001	0.725	0.124	<.0001
	Insurance (β_{c4})	−0.097	0.152	0.525			
	log(1+Income) (β_{c5})	0.042	0.016	0.008	0.040	0.014	0.004
	Wave (β_{c6})	0.131	0.025	<.0001	0.103	0.017	<.0001
	Heilongjiang (β_{c7})	0.212	0.309	0.492			
	Jiangsu (β_{c8})	−0.038	0.233	0.869			
	Shandong (β_{c9})	−0.461	0.288	0.110			
	Henan (β_{c10})	−0.337	0.263	0.201			
	Hubei (β_{c11})	0.172	0.246	0.484			
	Hunan (β_{c12})	0.367	0.254	0.149			
	Guangxi (β_{c13})	−0.176	0.254	0.488			
	Guizhou (β_{c14})	−0.516	0.247	0.037	−0.456	0.192	0.018
Heteroscedasticity	Intercept (δ_{c0})	−0.628	0.236	0.008	−0.532	0.063	<.0001
	Rural (δ_{c1})	−0.116	0.090	0.195			
	Age (δ_{c2})	0.002	0.003	0.409			
	Disease (δ_{c3})	0.109	0.088	0.216			
	Insurance (δ_{c4})	−0.068	0.115	0.554			
	log(1+Income) (δ_{c5})	0.013	0.010	0.180			
	Wave (δ_{c6})	0.023	0.020	0.244			
	Heilongjiang (δ_{c7})	−0.187	0.187	0.316			
	Jiangsu (δ_{c8})	−0.133	0.134	0.323			
	Shandong (δ_{c9})	−0.289	0.178	0.104			
	Henan (δ_{c10})	0.122	0.146	0.406			
	Hubei (δ_{c11})	−0.160	0.141	0.258			
	Hunan (δ_{c12})	−0.144	0.150	0.336			
	Guangxi (δ_{c13})	−0.006	0.139	0.965			
	Guizhou (δ_{c14})	−0.129	0.143	0.366			
	Kappa (κ)	2.545	0.269	<.0001	2.609	0.188	<.0001
Variance components	θ_A	0.509	0.034	<.0001	0.503	0.034	<.0001
	θ_B	1.561	0.048	<.0001	1.610	0.046	<.0001
	θ_{AB}	0.290	0.113	0.010	0.226	0.078	0.004

Table 6. Estimates (estimated standard errors) of marginal and incremental effects on the expected overall healthcare cost (Type II m2RGGMs) and on the expected positive healthcare cost (Type I m2RGGMs) at various combinations of rural status and disease status for the hypothetic individuals with an average age for the baseline province at the year of 2011 and, when it applies to the Type I m2RGGMs, for the hypothetic individuals with an average log income.

		Rural–disease combination		
		x_{it1} (Rural) =	x_{it3} (Disease) =	Estimate (SE)
Type II m2RGGMs	$\hat{\pi}_1(x_{it}, \hat{\vartheta})$	–	0	−31.989(14.181)
		–	1	−138.404(60.169)
	$\hat{\eta}_2(x_{it}, \hat{\vartheta})$	0	0	2.839(0.746)
		0	1	12.282(2.957)
		1	0	1.957(0.469)
		1	1	8.465(1.894)
	$\hat{\pi}_3(x_{it}, \hat{\vartheta})$	0	–	342.411(66.801)
		1	–	235.995(44.540)
Type I m2RGGMs	$\hat{\pi}_1(x_{it}, \hat{\vartheta})$	–	0	−204.576(96.459)
		–	1	−469.910(216.537)
	$\hat{\eta}_2(x_{it}, \hat{\vartheta})$	0	0	9.455(3.892)
		0	1	21.717(8.397)
		1	0	6.848(2.699)
		1	1	15.731(5.873)
	$\hat{\pi}_3(x_{it}, \hat{\vartheta})$	0	–	962.525(218.928)
		1	–	697.192(159.810)
	$\hat{\eta}_5(x_{it}, \hat{\vartheta})$	0	0	21.551(10.809)
		0	1	49.503(24.442)
		1	0	15.610(8.421)
		1	1	35.857(19.216)

The hyphens "–" indicate that it is not required to specify the covariate values when the corresponding marginal effects are estimated.

the expected positive healthcare cost. These effects are estimated under the reduced model in Table 3 for the hypothetic individuals whose age is equal to the sample average of the age variable and whose log income is equal to the sample average (i.e., $x_{it5} = 2.993$). Table 6 reports the estimates and the estimated SEs of the effects at various combinations of rural status and disease status for the baseline province at the year of 2011. The interpretations of these effects are similar to those in Table 4, but with respect to the expected positive healthcare expenditure. We also examine the average incremental effect of rural status $\bar{\pi}_1(\vartheta)$, the average marginal effect of age $\bar{\eta}_2(\vartheta)$, and the average incremental effect of disease status $\bar{\pi}_3(\vartheta)$ on the expected positive healthcare cost. The estimated average marginal effect of age and log income on the expected positive healthcare cost is 8.098 (3.293) and 18.459 (9.551), respectively. The estimated average incremental effects of rural status and disease status are −211.443 (96.138) and 552.109 (116.945), respectively. The estimated average semi-elasticities $\bar{s}_2(\vartheta)$ for age and $\bar{s}_5(\vartheta)$ for log income are instantly given by the corresponding estimated coefficients 0.013 (0.005) and 0.029 (0.014), respectively.

6 Discussion

In this article, we propose the Type I and Type II m2RGGms for modeling longitudinally observed healthcare costs and medical expenditures. We subsequently derive the estimates and variance estimates of various marginal effects of a covariate with respect to the expected overall and positive healthcare costs. Type I and Type II m2RGGms are joint random-effects models, in which the two modeling components are correlated thorough latent random effects. A major limitation of the proposed Type I and Type II m2RGGMs is that the computational complexity of maximizing the full likelihood to obtain parameter estimates increases dramatically if the total number of random effects in the two components of the m2RGGMs is increased. This is caused by the exponentially increased dimension of the integral resided in the full likelihood that I have to be numerically evaluated during the optimization process. However, this limitation is an inherent attribute as a random-effects model. Therefore, this is not unique in the proposed models and is true in most of the random-effects models. In fact, the proposed m2RGGMs retain the same level of computational complexity as the c2RGGMs in Liu et al.[10] In the statistical literature, the approaches that can resolve this issue have been well documented. One option is to find maximum likelihood estimates using an approximate Fisher scoring procedure based on high-order Laplace approximations as in Olsen and Schafer.[4] Other alternatives include developing a Monte Carlo expectation-maximization algorithm for maximizing the full likelihood[24,25] or adopting a pseudo-likelihood approach for parameter estimation.[26] The development and discussion on these three methods are beyond the scope of this article and can be pursued in the future research.

Direct marginal inference on the presence of healthcare utilization can be conducted thorough the marginalization in Part (I) of the Type I and Type II m2RGGMs. The corresponding marginal and incremental effects and their estimates can be similarly derived as in Section 3. It is not trivial to develop a marginalized model for the two-part random-effects model with an alternative distribution for $Y_{it}|Y_{it} > 0$ in Part (II) (see Liu et al.[27]). This has been identified as future research and will be reported in a separated manuscript. In addition, it is valuable as a topic of future research to compare the Type I and Type II m2RGGMs with the marginal models using the generalized estimating equation approach and other types of marginalized models that can provide direct marginal inference to healthcare cost panel data.

The development of the m2RGGMs in this article has connection with the work published in Su et al.[28] and Tom et al.,[29] but there is clear distinction between their work and ours. Su et al.[28] and Tom et al.[29] aimed at the marginal inference only on the probability of observing a positive cost in Part (I) of a two-part random-effects model and ignored Part (II). We instead propose to marginalize both Part (I) and Part (II), with particular emphasis on the marginal inference on the expected amount of overall healthcare costs. Also, Smith et al.[30] shares the same vision with our work as both of us plan to achieve marginal inference on overall healthcare costs. However, the concept ‘‘marginal inference’’ is used to highlight the fact that the regression models are not conditioning on *either* other response variables *or* unobserved random effects, and consequently can make direct inference on the effect of a covariate on marginal means.[31] The model proposed by Smith et al.[30] is still a conditional model (please refer to Diggle et al.[31] for connection and distinction between conditional models and marginal models) and therefore, as we discussed above, direct marginal inference cannot be achieved (see Liu et al.[10]). In addition, the m2RGGMs proposed here are built upon marginalization of the random-effects generalized Gamma models in Part (II) of the two-part models. This is beyond and superior to the lognormal models in Su et al.[28] and Smith et al.[30]

Acknowledgements

We sincerely thank two anonymous reviewers and the editors for their valuable comments, which had substantially improved this article. The views expressed in this article are those of the authors and do not represent the views of US Food and Drug Administration.

Authors' note

This research uses data from China Health and Nutrition Survey (CHNS).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Dr Wei Liu's research was partially supported by National Natural Science Foundation of China (Project 11601106). We thank the National Institute for Nutrition and Health, China Center for Disease Control and Prevention, Carolina Population Center (P2C HD050924, T32 HD007168), the University of North Carolina at Chapel Hill, the NIH (R01-HD30880, DK056350, R24 HD050924, and R01-HD38700) and the NIH Fogarty International Center (D43 TW009077, D43 TW007709) for financial support for the CHNS data collection and analysis files from 1989 to 2015 and future surveys, and the China-Japan Friendship Hospital, Ministry of Health for support for CHNS 2009, Chinese National Human Genome Center at Shanghai since 2009, and Beijing Municipal Center for Disease Prevention and Control since 2011.

References

1. Manning WG, Morris CN, Newhouse JP, et al. A two-part model of the demand for medical care: preliminary results from the health insurance study. In: van der Gaag J and Perlman M (eds) *Health, economics, and health economics*. Amsterdam, North Holland, 1981, pp.103–123.
2. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J Health Econ* 1998; **17**: 247–281.
3. Blough DK, Madden CW and Hornbrook MC. Modeling risk using generalized linear models. *J Health Econ* 1999; **18**: 153–171.
4. Olsen MK and Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. *J Am Stat Assoc* 2001; **96**: 730–745.
5. Tooze JA, Grunwald GK and Jones RH. Analysis of repeated measures data with clumping at zero. *Stat Meth Med Res* 2002; **11**: 341–355.
6. Duan N, Manning WG, Morris C, et al. A comparison of alternative models for the demand for medical care. *J Bus Econ Stat* 1983; **1**: 115–126.
7. Zhou XH, Stroupe KT and Tierney WM. Regression analysis of health care charges with heteroscedasticity. *J Roy Stat Soc Series C* 2001; **50**: 303–312.
8. Zhou XH, Lin H and Johnson E. Non-parametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *J Roy Stat Soc Series B* 2008; **70**: 1029–1047.
9. Manning WG, Basu A and Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. *J Health Econ* 2005; **24**: 465–488.
10. Liu L, Strawderman RL, Cowen ME, et al. A flexible two-part random effects model for correlated medical costs. *J Health Econ* 2010; **29**: 110–123.
11. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* 1999; **55**: 688–698.
12. Heagerty PJ and Zeger SL. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Stat Sci* 2000; **15**: 1–26.
13. Huang F and Gan L. The impacts of China's urban employee basic medical insurance on healthcare expenditures and health outcomes. *Health Econ* 2017; **26**: 149–163.
14. Wooldridge JM. *Econometric analysis of cross section and panel data*. 2nd ed. Cambridge: MIT Press, 2010.
15. Basu A and Rathouz PJ. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* 2005; **6**: 93–109.
16. Wooldridge JM. Unobserved heterogeneity and estimation of average partial effects. In: Andrews DWK and Stock JH (eds) *Identification and inference for econometric models: essays in honor of Thomas Rothenberg*. Cambridge: Cambridge University Press, 2005, pp.27–55.
17. Fernández-Val I. Fixed effects estimation of structural parameters and marginal effects in panel probit models. *J Econ* 2009; **150**: 71–85.
18. Shiu JL and Hu Y. Identification and estimation of nonlinear dynamic panel data models with unobserved covariates. *J Econ* 2013; **175**: 116–131.
19. Lei X and Lin W. The new cooperative medical scheme in rural China: does more coverage mean more service and better health? *Health Econ* 2009; **18**: S25–S46.
20. National Bureau of Statistics of China. *China statistical yearbook*. Beijing: China Statistics Press, 2004–2011.
21. Yip W, Hsiao WC, Chen W, et al. Early appraisal of China's huge and complex health-care reforms. *Lancet* 2012; **379**: 833–842.
22. Wagstaff A, Lindelow M, Gao J, et al. Extending health insurance to the rural population: an impact evaluation of China's new cooperative medical scheme. *J Health Econ* 2009; **28**: 1–19.
23. Gong G and Samaniego FJ. Pseudo maximum likelihood estimation: theory and applications. *Ann Stat* 1981; **9**: 861–869.
24. McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *J Am Stat Assoc* 1997; **92**: 162–170.
25. Booth JG and Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J Roy Stat Soc Series B* 1999; **61**: 265–285.
26. Varin C, Reid N and Firth D. An overview of composite likelihood methods. *Stat Sin* 2011; **21**: 5–42.
27. Liu L, Strawderman RL, Johnson BA, et al. Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. *Stat Meth Med Res* 2016; **25**: 133–152.
28. Su L, Tom BD and Farewell VT. A likelihood-based two-part marginal model for longitudinal semicontinuous data. *Stat Meth Med Res* 2015; **24**: 194–205.
29. Tom BD, Su L and Farewell VT. A corrected formulation for marginal inference derived from two-part mixed models for longitudinal semi-continuous data. *Stat Meth Med Res* 2013; **25**: 2014–2020.
30. Smith VA, Neelon B, Preisser JS, et al. A marginalized two-part model for longitudinal semicontinuous data. *Stat Meth Med Res* 2013; **25**: 1949–1968.
31. Diggle P, Heagerty P, Liang KY, et al. *Analysis of longitudinal data*. Oxford: Oxford University Press, 2002.