Aust. N. Z. J. Stat. 2020 doi: 10.1111/j.1467-842X.XXX

A shared parameter mixture model for longitudinal income data with missing responses and zero rounding

Francis K. C. Hui [1*], and Howard D. Bondell[2]

Australian National University and The University of Melbourne

Summary

The analysis of longitudinal income data is often made challenging for several reasons. For example, in a national Australian survey on income over time, a non-negligible proportion of responses are missing, and it is believed the missingness mechanism is non-ignorable. Also, there are a large number of reported zero incomes, some of which may be true zeros (corresponding to individuals who legitimately do not earn an income), while some may be false zeros (corresponding to individuals choosing to round their income to zero). We propose a new shared parameter mixture (SPM) model for analysing semicontinuous longitudinal income data, which addresses the two challenges of income non-response and zero rounding. This is accomplished by jointly modelling an individual's underlying income together with the probability of missingness and rounding to zero, where both probabilities are permitted to vary in a smooth manner with their underlying non-zero income. Applying the SPM model to the Australian income survey reveals that on average, older female individuals and individuals with a long term health condition are considerably less likely to earn an income, while income tended to be highest for male individuals on fixed–term/permanent job contracts between ages 50 to 60. Furthermore there is evidence of both zero rounding, and conditional on the assumed missingness mechanism, individuals with incomes at the higher and lower ends are more likely to not report their income.

Key words: income; missing not at random; mixed models; non-ignorable; zero inflation

1. Introduction

Modelling how income distribution changes over time with an individual's physical and social circumstances is of major interest in economics, with important implications for understanding the relationship between income and public health issues such as poverty, and

*Author to whom correspondence should be addressed.
[1] Research School of Finance, Actuarial Studies & Statistics, Australian National University, Acton, ACT 2601, Australia
[2] School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC 3010, Australia
Email: francis.hui@anu.edu.au

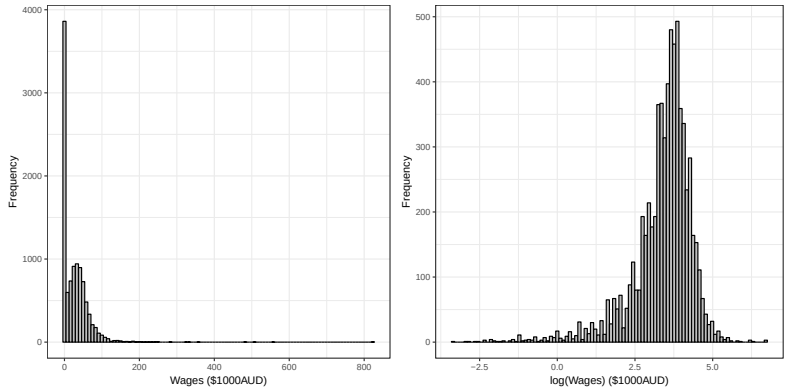
Acknowledgment. Both FKCH and HDB were supported Australian Research Council Discovery projects. Thanks to Nicole Watson for useful discussions relating to the HILDA survey. The HILDA Survey was initiated and is funded by the Australian Government Department of Social Services and is managed by the Melbourne Institute of Applied Economic and Social Research. The findings and views reported in this paper, however, are those of the authors.

Prepared using anzsauth.cls *[Version: 2016/19/05 v1.02]*

poor physical and mental health (e.g., Bechtel, Lordan & Rao 2012). This article is motivated by the Household Income and Labour Dynamics in Australia (HILDA) survey (Watson & Wooden 2012), a national panel survey collected annually in Australia since 2001. One of the main goals of the HILDA survey is to achieve a greater understanding of how individual earnings have evolved over time in Australia, along with factors that drive income distribution and its evolution.

As is the case more generally, modelling income data in the HILDA survey is made challenging for two main reasons. First, a proportion of the income responses (approximately 7.60%) are missing, and it is often more realistic to assume that non-response is missing not at random (Rubin 1976) as individuals with very low or very high income tend to refuse to report their income (Riphahn & Serfling 2005; Schraepler 2006). While the proportion of data missing is not as large as in other national surveys on income (e.g., the British Household Panel Survey, where it exceeds 15%, Jenkins 2010), it remains true that with such non-response we can not ignore the missing data mechanism because it invalidates many standard methods of inference on longitudinal data (Little & Rubin 2014). In the context of longitudinal income data, there has been some research on handling non-ignorable responses (e.g., the pattern mixture model of Giusti & Little 2011), although the vast majority of developments have instead focussed on assuming income is missing at random and thus ignorable missingness, largely driven by the easier means of analyses, and comparing various imputation methods (see Headey & Wooden 2004; Watson & Starick 2011, for examples with the HILDA survey). A second major challenge is that there are a substantial number of incomes (approximately 32.48%) reported as zero, as exemplified in Figure 1. In other disciplines, such data comprising of one (or more) spikes at specific values along with a continuous distribution are known are semicontinuous data (Bohning & Alfo 2016; Liu et al. 2019). While a proportion of these reported zero incomes may correspond to true zeros e.g., the individual is retired, some may be heaped or rounded zeros arising from an individual with very low but non-zero income choosing to round their income to zero (see for instance Hanisch 2005, for known evidence of this phenomenon). If such a rounding mechanism is not accounted for i.e., all the zeros are assumed to be true zeros, then this can also invalidate inference on the likelihood of a person having a non-zero response. Recent work by Zinn & Wrbach (2016) and Gross & Rendtel (2016) among others have proposed heuristic approaches to first identify values to which incomes tend to be rounded, and then use kernel-type heaping functions to quantify the probability of an income value being rounded. However, both articles were solely focused on modelling the rounding process, while we are interested in modelling potential rounding to zero as part of an overall analysis for how income varies with an individual's attributes over time.

Figure 1. Histograms of all reported income data (left) and $\log(\text{income})$ for reported non-zero income data only (right) from the HILDA survey. In the left panel, we can see a substantial proportion of incomes reported equal to zero, while the distribution of non-zero incomes in the right panel is unimodal and presents evidence of rounding at non-zero income values.



In this article, we propose a shared parameter mixture (SPM) model for analysing semicontinuous longitudinal income data, which explicitly addresses the challenges of non-ignorable responses and zero rounding in the HILDA survey. To our knowledge, our article is the first in the applied statistics literature that attempts to overcome both challenges simultaneously in a single model. This is despite the fact that both missing responses and rounding tend to occur at the same time in many income surveys. The SPM model is formulated hierarchically by first constructing a latent process relating an individual's true mean income to their covariate values. This takes the form of a mixture model, with one mixture component being a point mass at zero used to model the probability of an individual earning an income, and the second mixture component being a positive continuous distribution used to model the mean non-zero income over time, conditional on earning an income. For modelling semicontinuous responses comprising a point mass at zero and positive continuous data, mixture models are a popular although by no means the only approach; see the special issue edited by Bohning & Alfo (2016) as well as the recent review article by Liu et al. (2019). More importantly, we believe this mixture model is appropriate for studying the true underlying income distribution in the HILDA survey, as it ensures that the two processes of whether an individual actually earns an income and then how much they earn are modelled explicitly, with potentially differing covariates driving the two mean structures characterising these processes; see also Bacci et al. (2019).

Given the latent income process, the SPM model then builds in two components which account for potential non-ignorable missingness and zero rounding. For the former, we propose a shared parameter logistic regression model, such that the probability of missingness

depends on the value of their underlying mean non-zero income as well as other covariates. This approach is similar to shared parameter regression models, where random effects are used to drive both, and thus induce a dependence between, the responses and the missing data mechanism (Chapter 19, Fitzmaurice et al. 2008). However, a key difference is that rather than sharing the random effects, we share the conditional mean in the non-zero component of the underlying mixture model, such that the probability of non-response is allowed to vary in a smooth manner with underlying non-zero income. Apart from being more parsimonious, this approach also improves interpretability as we can differentiate between direct effects and indirect effects (acting via income) of covariates on the likelihood of missingness. Indeed, previous studies have shown that, aside from being more likely for lower and higher income earners, income non-response may also be directly affected by personal attributes such as age, which also affect the level of income (Riphahn & Serfling 2005). As a side note, we point out that approaches where the linear predictor (or function of it) is included in the model for missingness are sometimes referred to as 'joint models' (Wulfsohn & Tsiatis 1997). However in this article, we will use the terminology 'shared parameter models' as is done in Rizopoulos, Verbeke & Molenberghs (2008), Tsonaka, Verbeke & Lesaffre (2009) and Creemers et al. (2011), among others. Conditional on the responses not being missing, we then propose a second logistic regression model to account for potential zero rounding. This component of the SPM model may be also regarded as a shared parameter model, as the likelihood of rounding depends on the true income and thus indirectly on covariates driving income.

We estimate and perform inference on the SPM model using Bayesian v Chain Monte Carlo (MCMC) estimation via JAGS (Just Another Gibbs Sampler, Plummer et al. 2003). Simulations demonstrate that the SPM model can successfully capture potential non-ignorable missingness and zero rounding in the income responses, while models that do not explicitly account for such processes may produce incorrect inference on the predictors driving the income distribution and its evolution. Applying the SPM model to the HILDA survey produced several interesting although not entirely surprising results: 1) older female individuals and individuals with a long term health condition are comparably less likely to earn an income; 2) males between ages 50–60 on a permanent or fixed-term job contract tend to earn the most income. Furthermore, income distribution depends on age in a nonlinear manner; 3) there is a strong association between the probability of zero rounding and a person's true non-zero income. A sensitivity analysis also suggests that many of the parameter estimates are sensitive to the missing data assumptions. We provide template R code for fitting the SPM model, as well to reproduce the simulation and application in Supplementary Material C.

2. The shared parameter mixture model

Consider a longitudinal dataset comprising n independent individuals, where for individual $i = 1, \ldots, n$ we let $y_{ij,\text{obs}}$ denote their reported income at time point $j = 1, \ldots, n_i$. In the motivating HILDA survey, there are $n = 1,693$ individuals who have reported incomes ranging from $n_i = 2$ to 10 time points, leading to a total of 10,650 observations. Along with the responses, we observe a number of explanatory variables e.g., gender and age. For the HILDA survey, all covariate values in the dataset are observed, while some of the $y_{ij,\text{obs}}$'s are missing. In what follows, we shall use $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ to denote covariates to be included as fixed effects and random effects, respectively. Unless stated otherwise, both covariate vectors are assumed to contain an intercept term as their first elements. Moreover, as in common in longitudinal mixed models, the elements of these vectors tend to overlap since random slopes are generally also included as fixed slopes (Hui, Mller & Welsh 2017). We formulate the SPM model by first describing the latent process governing an individual's true income, before discussing the components mapping this latent process to the observed income.

We propose a two-component mixture model to quantify the latent process for an individual's true income, denoted here as $y_{ij,\text{true}}$. Let g_{ij} be an indicator of whether individual i earns an income at time point j, such that $g_{ij} = 1$, if $y_{ij,\text{true}} > 0$, and 0 otherwise. Note g_{ij} is not fully observed, as we shall discuss later. Then we apply a logistic regression to model the probability that an individual earns an income. Conditional on a vector of random effects $\boldsymbol{b}_{i,\text{earn}}$, we assume the indicators g_{ij} are independent Bernoulli random variables,

$$f(g_{ij}|\boldsymbol{b}_{i,\text{earn}}) = \text{Bern}(g_{ij}; \pi_{ij}); \quad \text{logit}(\pi_{ij}) = \boldsymbol{x}_{ij,\text{earn}}^\top \boldsymbol{\beta}_{\text{earn}} + \boldsymbol{z}_{ij,\text{earn}}^\top \boldsymbol{b}_{i,\text{earn}}, \tag{1}$$

where $f(\cdot|\cdot)$ is used to denote conditional distributions, $\boldsymbol{\beta}_{\text{earn}}$ are the fixed effects coefficients and $\text{Bern}(\cdot; \mu)$ denotes the Bernoulli distribution with probability of success μ. The random effects are assumed to be drawn from a multivariate normal distribution, such that $f(\boldsymbol{b}_{i,\text{earn}}) = \mathcal{N}(\boldsymbol{b}_{i,\text{earn}}; \boldsymbol{0}, \boldsymbol{\Sigma}_{\text{earn}})$ for some unstructured covariance matrix $\boldsymbol{\Sigma}_{\text{earn}}$, where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Second, conditional on earning an income i.e., $g_{ij} = 1$, we model the individual's true non-zero income via a second regression model,

$$\begin{aligned} f(y_{ij,\text{true}}|g_{ij} = 1, \boldsymbol{b}_{i,\text{nonzero}}) &= \mathcal{F}_+(y_{ij,\text{true}}; \eta_{ij}, \phi); \\ \eta_{ij} &= \boldsymbol{x}_{ij,\text{nonzero}}^\top \boldsymbol{\beta}_{\text{nonzero}} + \boldsymbol{z}_{ij,\text{nonzero}}^\top \boldsymbol{b}_{i,\text{nonzero}}, \end{aligned} \tag{2}$$

where $\mathcal{F}_+(\cdot; \eta, \phi)$ denotes a distribution defined on the positive real line and characterised by parameters (η, ϕ) as defined below, and $f(\boldsymbol{b}_{i,\text{nonzero}}) = \mathcal{N}(\boldsymbol{b}_{i,\text{nonzero}}; \boldsymbol{0}, \boldsymbol{\Sigma}_{\text{nonzero}})$. We consider three choices for $\mathcal{F}_+(\cdot; \eta, \phi)$, although we acknowledge more complex distributions for

income are possible (e.g., Caldern-Ojeda, Azpitarte & Gmez-Dniz 2016): 1) a gamma distribution with mean $\mu = \exp(\eta)$ and shape parameter ϕ; 2) an exponential distribution with mean $\mu = \exp(\eta)$. This is a special case of the gamma distribution with shape parameter $\phi = 1$; 3) a lognormal distribution with location η and variance ϕ parameters on the log scale, such that the mean of the distribution is equal to $\mu = \exp(\eta + 2^{-1}\phi)$. In all three cases, we note the use of an exponential link function to reflect the notion that wealth is driven by a series of multiplicative processes. Let $\boldsymbol{\Psi}_{\text{earn}} = (\boldsymbol{\beta}_{\text{earn}}, \text{vech}(\boldsymbol{\Sigma}_{\text{earn}}))$ and $\boldsymbol{\Psi}_{\text{nonzero}} = (\boldsymbol{\beta}_{\text{nonzero}}, \text{vech}(\boldsymbol{\Sigma}_{\text{nonzero}}), \phi)$ denote the parameters describing the probability of earning an income and the distribution of non-zero incomes respectively. Furthermore, by marginalising (2) with respect to g_{ij} and (1), the formulation of the mixture model becomes clear as we see that $f(y_{ij,\text{true}}|\boldsymbol{b}_{i,\text{true}}, \boldsymbol{b}_{i,\text{nonzero}}) = (1 - \pi_{ij})\mathbb{I}_0 + \pi_{ij}\mathcal{F}_+(y_{ij,\text{true}}; \eta_{ij}, \phi)$, where $\mathbb{I}_0$ is the Dirac delta function. The distribution of true income is thus a mixture of a spike at zero and a positive continuous distribution, and as reviewed in Section 1 such a model is commonly applied for analysing semicontinuous data.

We next formulate a model for the missing responses. Let r_{ij} denote the missing data indicators, such that $r_{ij} = 1$ implies $y_{ij,\text{obs}}$ is missing at time point j for individual i, and $r_{ij} = 0$ otherwise. Also, let $\mu_{ij} = \text{E}(y_{ij,\text{true}}|g_{ij} = 1, \boldsymbol{b}_{i,\text{nonzero}})$ denote the conditional mean non-zero income based on (2), e.g., $\mu_{ij} = \exp(\eta_{ij} + 2^{-1}\phi)$ if $y_{ij,\text{true}}$ is assumed to follow a lognormal distribution. To allow for income missingness mechanism being potentially non-ignorable, we allow the probability of missingness to vary as a function of μ_{ij} (conditional on the individual earning an income) as well as with a set of fixed and random effect covariates. That is, given a vector of random effects $\boldsymbol{b}_{i,\text{miss}}$, we assume the r_{ij}'s are independent Bernoulli random variables,

$$\begin{aligned} f(r_{ij}|g_{ij}, \boldsymbol{b}_{i,\text{nonzero}}, \boldsymbol{b}_{i,\text{miss}}) &= \text{Bern}(r_{ij}; \rho_{ij}); \\ \text{logit}(\rho_{ij}) &= \boldsymbol{x}_{ij,\text{miss}}^\top \boldsymbol{\beta}_{\text{miss}} + \boldsymbol{z}_{ij,\text{miss}}^\top \boldsymbol{b}_{i,\text{miss}} + g_{ij} s_{\text{miss}}(\mu_{ij}), \end{aligned} \tag{3}$$

where $s_{\text{miss}}(\cdot) : \mathbb{R} \to \mathbb{R}$ is some smooth function satisfying $s_{\text{miss}}(0) = 0$, and $f(\boldsymbol{b}_{i,\text{miss}}) = \mathcal{N}(\boldsymbol{b}_{i,\text{miss}}; \boldsymbol{0}, \boldsymbol{\Sigma}_{\text{miss}})$. Conditional on earning an income ($g_{ij} = 1$), we see that due to shared random effects between μ_{ij} and ρ_{ij}, there is a resulting marginal dependence between the missing data indicators and the missing observations themselves. Consider an individual who always earn an income, such that $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{in_i}) = \boldsymbol{1}$. If we let $\boldsymbol{y}_{i,\text{true}}$ and $\boldsymbol{r}_i$ denote the n_i-vectors of true income and missing response indicators for individual i, respectively, then

$$f(\boldsymbol{r}_i|\boldsymbol{y}_{i,\text{true}}) = \int \prod_{j=1}^{n_i} f(r_{ij}|\boldsymbol{b}_{i,\text{nonzero}}, \boldsymbol{b}_{i,\text{miss}}, \boldsymbol{y}_{i,\text{true}}) f(\boldsymbol{b}_{i,\text{nonzero}}|\boldsymbol{y}_{i,\text{true}}) f(\boldsymbol{b}_{i,\text{miss}}) d\boldsymbol{b}_{i,\text{nonzero}} d\boldsymbol{b}_{i,\text{miss}},$$

where

$$f(r_{ij}|\boldsymbol{b}_{i,\text{nonzero}}, \boldsymbol{b}_{i,\text{miss}}, \boldsymbol{y}_{i,\text{true}}) = \text{Bern}\left[r_{ij}; \text{logit}^{-1}\left\{\boldsymbol{x}_{ij,\text{miss}}^{\top}\boldsymbol{\beta}_{\text{miss}} + \boldsymbol{z}_{ij,\text{miss}}^{\top}\boldsymbol{b}_{i,\text{miss}} + s_{\text{miss}}(\mu_{ij})\right\}\right].$$

Therefore the dependence of the missing indicators on underlying income is made explicit through the dependence on the posterior distribution of the random effects, $f(\boldsymbol{b}_{i,\text{nonzero}}|\boldsymbol{y}_{i,\text{true}})$. An analogous dependence result can be obtained for individuals who vary between earning and not earning an income i.e, $\boldsymbol{g}_i$ contains both zeros and ones.

Unlike standard applications of shared parameter models, where the the random effects $\boldsymbol{b}_{i,\text{nonzero}}$ are included directly in the linear predictor (e.g., Fitzmaurice et al. 2008), the probability of missingness in the SPM model is dependent on the mean non-zero income μ_{ij}. The motivation behind this (at least with the current parameterisation) is to distinguish between the indirect effects of covariates on missingness occurring through the responses, and the direct effects of covariates on the probability of missingness. For example, if an individual earns a very large income then they may choose not to report it due to its large value, and since income typically has a significant relationship with age (as is the case in the HILDA survey; see Section 5), then in such case age has an *indirect* effect on the probability of missingness through μ_{ij}. On the other hand, the probability of missingness may also be directly related to age e.g., as respondents tend to become more familiar with the survey over time and thus are more comfortable with reporting their income, in which case age has a direct effect as captured by the $\boldsymbol{\beta}_{\text{miss}}$ and $\boldsymbol{b}_{i,\text{miss}}$. Note also that by allowing the dependence to be on μ_{ij} rather than on $\boldsymbol{b}_{i,\text{nonzero}}$, we require only a small number of parameters to capture the non-ignorable missingness in (3), depending on the choice of $s_{\text{miss}}(\cdot)$. Given the large number of parameters already present in the SPM model, the ability to capture a non-ignorable missingness mechanism parsimoniously yet flexibly is an advantage.

It is worth discussing in more detail how (3) allows for and transitions between the cases of $g_{ij} = 0$ and $g_{ij} = 1$ in a smooth manner. If $g_{ij} = 0$, then an individual does not earn income and the probability of non-response can be considered as being depending upon the underlying, zero income. If $g_{ij} = 1$, then an individual earns income and the probability of non-response is allowed to vary with the underlying income. However, the former may also be interpreted as a limiting case of the latter when $\mu_{ij} \to 0$, that is, we suppose an individual earns an income but consider how the probability of non-response varies as their income tends to zero. To allow this transition to occur in a smooth manner, we constrain $s_{\text{miss}}(0) = 0$. Moreover, we see that the covariates $(\boldsymbol{x}_{ij,\text{miss}}, \boldsymbol{z}_{ij,\text{miss}})$ in (3) can not only be interpreted as direct effects on the probability of missingness (as discussed), but are in fact the only effects in the limiting case of $\mu_{ij} = 0$.

Regarding the choice of $s_{\text{miss}}(\cdot)$, while it is possible to consider more data-driven smoothing functions, for ease of interpretation (and likely the inability to estimate such terms effectively without much missing data) we choose at most to use a quadratic form $s_{\text{miss}}(\mu) = \mu\theta_1 + \mu^2\theta_2$, where $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The quadratic form also means the probability of missingness may vary nonlinearly as a function of the mean response, even if the mean itself only varies linearly in time. For instance, an individual's income may increase monotonically over time, but by constructing $s_{\text{miss}}(\cdot)$ as a quadratic function then the probability of non-response is (say) higher at earlier and later time points in the survey, when the individual has relatively low and high income, respectively. Also, at a single time point it allows for low income and high income earners to have higher propensity of non-response relative to average income earners. Let $\boldsymbol{\Psi}_{\text{miss}} = (\boldsymbol{\beta}_{\text{miss}}, \text{vech}(\boldsymbol{\Sigma}_{\text{miss}}), \boldsymbol{\theta})$ denote the parameter vector corresponding to the component of the SPM model characterising the missing data process. In our application to the HILDA survey, we conduct a sensitivity analysis by comparing different choices for $s_{\text{miss}}(\cdot)$, including the case of $s_{\text{miss}}(\cdot) = 0$, and its impact on the parameter estimates in the SPM model.

We remark that making use of random effects is by no means the only approach for modelling the serial correlation between the indicators of missing income; alternative approaches such as assuming a Markovian process to account for dependence between neighbouring r_{ij} (and indeed the g_{ij} for that matter) are possible; see Chapter 18, Fitzmaurice et al. (2008). Our main motivation for using random effects to account for the temporal correlation is precisely because it provides a relatively straightforward mechanism to additionally induce correlation between the missingness indicators and the observed income, and indeed also with indicator of zero rounding as we see below; see also Bacci & Bartolucci (2015).

The last component of the SPM model involves constructing a logistic regression model for the probability that an individual rounds their true income to zero, conditional on earning an income. Let $h_{ij} = 1$ if $y_{ij,\text{true}}$ is rounded to zero and 0 otherwise. As with g_{ij}, note that h_{ij} is not fully observed. We model the h_{ij}'s as conditionally independent Bernoulli random variables,

$$f(h_{ij}|g_{ij} = 1, \boldsymbol{b}_{i,\text{nonzero}}) = \text{Bern}(h_{ij}; \nu_{ij}); \quad \text{logit}(\nu_{ij}) = s_{\text{round}}(\mu_{ij}), \tag{4}$$

such that ν_{ij} is the probability of rounding to zero, and $s_{\text{round}}(\cdot) : \mathbb{R} \to \mathbb{R}$ is a smooth function. Analogous to (3), we allow the probability of zero rounding to vary with the conditional mean non-zero income μ_{ij}. Moreover as this is now shared between the models governing $y_{ij,\text{true}}, r_{ij}$ and h_{ij}, it induces a correlation between all three observed quantities and causing them to co-vary with each other. Note we choose not to include covariates in the linear

predictor in (4), as we argue that the propensity of an individual rounding their income to zero depends solely on the proximity of that income to zero (analogous to Zinn & Wrbach 2016; Gross & Rendtel 2016). Regarding the form of $s_{\text{round}}(\cdot)$, we only consider a simple linear form $s_{\text{round}}(\mu) = \tau_0 + \mu\tau_1$ where $\boldsymbol{\tau} = (\tau_0, \tau_1)$, with a negative value of τ_1 meaning that the probability of rounding to zero decreases as the true income increases. Let $\boldsymbol{\Psi}_{\text{round}} = \boldsymbol{\tau}$ denote the parameter vector corresponding to the component of the SPM model governing the zero rounding process.

It is important to acknowledge that we do not incorporate a mechanism to deal with rounding or heaping at non-zero income values, even though there is evidence of this in our motivating HILDA survey (Figure 1, right panel). Our reasoning behind this is that, in the given dataset, the fraction of tied data points at any one *non-zero* income value is at most 1.60% of the total number of observations (170 out of 10,650 observations occurring at $40,000 AUD), and as in typical practice, income data, although technically discrete, are modelled as approximately continuous. This contrasts to the heaping at zero, which accounts for approximately 32.5% of the total sample size. Of course, it is possible to apply data-driven and nonparametric methods to both select the rounding points and model the rounding process (e.g., Gross & Rendtel 2016). However we believe that, apart from both the estimation procedure being considerably more complicated as a result, inference on the non-zero income component of the SPM model will not be substantially affected by this.

2.1. Special cases

The SPM model includes two special cases of interest. First, if there is no missing data then (3) can be omitted with $r_{ij} = 1$ for all i and j, and the zero rounding process given by (2) becomes the sole process affecting $y_{ij,\text{true}}$. Second, if there is no zero rounding then (4) can be omitted with $h_{ij} = 1$ for all i and j. Note if there were no missing data also then it would be possible to treat the two mixture components defined by (1) and (2), separately. That is, estimation and inference for the logistic regression model with parameters $\boldsymbol{\Psi}_{\text{earn}}$ could be to done separately to the regression model for non-zero income with parameters $\boldsymbol{\Psi}_{\text{nonzero}}$. But with missing responses separate estimation is not possible due to the shared missing data mechanism driving both components of the mixture model. More broadly, in the full SPM model the effect of zero rounding is to induce a dependence between the two mixture components of (1) and (2) at the observation level. That is, observations from the non-zero income component can be moved into the zero income component, leading to cases of $y_{ij,\text{obs}} = 0$ even though $y_{ij,\text{true}} > 0$.

With longitudinal income data, neither simplification discussed above may be appropriate: there are almost always cases of non-response in income surveys as individuals

refuse to report their income (see also the sensitivity analysis in Supplementary Material B.4), while zero rounding and indeed rounding to other non-zero amounts is a common occurrence as seen in Figure 1. In Section 4, we show that if there is zero rounding and/or non-ignorable missingness and this is not accounted for in the analysis, then it can lead to erroneous inference for one or more components of the SPM model. Furthermore, in Section 5, we will see that the HILDA survey presents evidence of both features being present.

3. Estimation

The full formulation of the SPM model can be written as follows. For individual $i = 1, \ldots, n$ and time point $j = 1, \ldots, n_i$, denote the vector of all responses as $(y_{ij,\text{obs}}, r_{ij}, g_{ij}, h_{ij})$, i.e, the observed income, the missingness indicator, the indicator of whether the true income is positive, and the indicator of whether the true income has been rounded to zero. Apart from $y_{ij,\text{obs}}$, both g_{ij} and h_{ij} are only partly observed: if $y_{ij,\text{obs}} > 0$ then we know $g_{ij} = 1$ and $h_{ij} = 0$ i.e., the individual truly earns an income and chose not to round at that time point. If $y_{ij,\text{obs}} = 0$ then both g_{ij} and h_{ij} are missing since we do not know whether the individual actually did not earn an income, or did earn an income but chose to round their income to zero. Finally, if $r_{ij} = 1$ then $y_{ij,\text{obs}}$ is missing, and we also set g_{ij} and

h_{ij} to be missing. The SPM model is formulated as follows:

$$
\begin{aligned}
f(\boldsymbol{b}_{i,\text{earn}}) &= \mathcal{N}(\boldsymbol{b}_{i,\text{earn}}; \mathbf{0}, \boldsymbol{\Sigma}_{\text{earn}}) \\
f(\boldsymbol{b}_{i,\text{nonzero}}) &= \mathcal{N}(\boldsymbol{b}_{i,\text{nonzero}}; \mathbf{0}, \boldsymbol{\Sigma}_{\text{nonzero}}) \\
f(\boldsymbol{b}_{i,\text{miss}}) &= \mathcal{N}(\boldsymbol{b}_{i,\text{miss}}; \mathbf{0}, \boldsymbol{\Sigma}_{\text{miss}}) \\
f(g_{ij}|\boldsymbol{b}_{i,\text{earn}}) &= \text{Bern}(g_{ij}; \pi_{ij}); \\
&\quad \text{where } \text{logit}(\pi_{ij}) = \boldsymbol{x}_{ij,\text{earn}}^\top \boldsymbol{\beta}_{\text{earn}} + \boldsymbol{z}_{ij,\text{earn}}^\top \boldsymbol{b}_{i,\text{earn}}, \\
f(y_{ij,\text{true}} = 0|g_{ij} = 0) &= 1 \\
f(y_{ij,\text{true}}|g_{ij} = 1, \boldsymbol{b}_{i,\text{nonzero}}) &= \mathcal{F}_+(y_{ij,\text{true}}; \eta_{ij}, \phi); \\
&\quad \text{where } \eta_{ij} = \boldsymbol{x}_{ij,\text{nonzero}}^\top \boldsymbol{\beta}_{\text{nonzero}} + \boldsymbol{z}_{ij,\text{nonzero}}^\top \boldsymbol{b}_{i,\text{nonzero}}, \\
f(r_{ij}|g_{ij}, \boldsymbol{b}_{i,\text{nonzero}}, \boldsymbol{b}_{i,\text{miss}}) &= \text{Bern}(r_{ij}; \rho_{ij}); \\
&\quad \text{where } \text{logit}(\rho_{ij}) = \boldsymbol{x}_{ij,\text{miss}}^\top \boldsymbol{\beta}_{\text{miss}} + \boldsymbol{z}_{ij,\text{miss}}^\top \boldsymbol{b}_{i,\text{miss}} + g_{ij} s_{\text{miss}}(\mu_{ij}), \\
f(h_{ij}|g_{ij} = 1, \boldsymbol{b}_{i,\text{nonzero}}) &= \text{Bern}(h_{ij}; \nu_{ij}); \\
&\quad \text{where } \text{logit}(\nu_{ij}) = s_{\text{round}}(\mu_{ij}), \\
f(y_{ij,\text{obs}} = 0|g_{ij} = 0, r_{ij} = 0) &= f(y_{ij,\text{true}} = 0|g_{ij} = 0) = 1 \\
f(y_{ij,\text{obs}} = 0|g_{ij} = 1, r_{ij} = 0, h_{ij} = 1) &= 1 \\
f(y_{ij,\text{obs}}|g_{ij} = 1, r_{ij} = 0, h_{ij} = 0) &= f(y_{ij,\text{true}}|g_{ij} = 1, \boldsymbol{b}_{i,\text{nonzero}}),
\end{aligned}
$$

and $y_{ij,\text{obs}}$ is missing if $r_{ij} = 1$. To clarify the final three lines of the SPM model, note we have the following conditional distributions for reported income: 1) $f(y_{ij,\text{obs}} = 0|g_{ij} = 0, r_{ij} = 0) = f(y_{ij,\text{true}} = 0|g_{ij} = 0) = 1$, which has the interpretation that if an individual does not earn an income and choose to report it, then we observe a zero income; 2) $f(y_{ij,\text{obs}} = 0|g_{ij} = 1, r_{ij} = 0, h_{ij} = 1) = 1$, which has the interpretation that if an individual earns an income, chooses to report it but rounds to zero, then we observe a zero income; 3) $f(y_{ij,\text{obs}}|g_{ij} = 1, r_{ij} = 0, h_{ij} = 0) = f(y_{ij,\text{true}}|g_{ij} = 1, \boldsymbol{b}_{i,\text{nonzero}}) = \mathcal{F}_+(y_{ij,\text{true}}; \mu_{ij}, \phi)$, which means that if individual earns an income, chooses to report and does not round, then the distribution of the observed income is the same as the distribution of the true income $y_{ij,\text{true}}$.

The hierarchical nature of SPM model facilitates using Bayesian estimation via MCMC sampling, which we implemented using JAGS and the R package runjags (Denwood 2016) to jointly estimate all parameters (and random effects) in the model. As prior distributions for both our simulation study and application to the HILDA survey, for all the fixed effect coefficients $\boldsymbol{\beta}_{\text{earn}}, \boldsymbol{\beta}_{\text{nonzero}}, \boldsymbol{\beta}_{\text{miss}}$ and parameters $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ characterising $s_{\text{miss}}(\cdot)$ and $s_{\text{round}}(\cdot)$, respectively, we assigned independent normal priors to each element with mean zero and variance equal to ten, noting all covariates entered into the SPM model were standardised

to have mean zero and variance one. For the nuisance parameter ϕ in the gamma and lognormal distributions, we assigned a half-normal prior with scale parameter equal to ten. Our motivation for using such weakly informative priors for these parameters was to stabilise the MCMC sampling but also reflect the prior belief that these coefficients were unlikely to exceed the range ± 10 (e.g., Gelman et al. 2008). In both the simulations and application, we did test the use of more diffuse prior distributions e.g., setting variance/scale parameters to 50 instead of 10, and found that results tended to be similar overall but that convergence of the MCMC sampling took longer. Finally, for the three random effects covariance matrices $\boldsymbol{\Sigma}_{\text{earn}}$, $\boldsymbol{\Sigma}_{\text{nonzero}}$ and $\boldsymbol{\Sigma}_{\text{miss}}$, we assigned independent inverse Wishart priors with the degrees of freedom set equal to the dimension of the corresponding covariance matrix plus one, and the scale matrix set equal to the identity matrix in the case of the simulation study.

We ran three MCMC chains, with each chain consisting for 20,000 burn-in iterations followed by 50,000 additional samples with a thinning period of 25. We assessed convergence based on trace plots and the Gelman-Rubin convergence statistic (Gelman & Rubin 1992) and, if required (although this was rarely needed), extended the MCMC sampling until the convergence statistic was below 1.1 for all parameters in the model. Without extension, this lead to a total of 6,000 MCMC samples for analysis.

4. Simulation study

We conducted an empirical study to assess the performance of the SPM model, and to illustrate how the failure to account for possible zero rounding and/or non-ignorable missingness can lead to erroneous inference on various aspects of the longitudinal income data. We used the SPM model fitted to the HILDA survey in Section 5 as the basis for constructing a true SPM model as follows. First, we included only fixed effect covariates for which the 95% highest posterior density interval of the corresponding regression coefficient from the fitted SPM model did not contain zero. Next, we considered only a simple linear relationship for the fixed effect of age in the simulation, rather than the non-linear relationship modelled based on spline functions in the real application. This allowed us to more straightforwardly assess the performance of competing models in terms of inference on the effect of age. Based on these two modifications, this led to six coefficients in $\boldsymbol{\beta}_{\text{earn}}$, five coefficients in $\boldsymbol{\beta}_{\text{nonzero}}$, and four coefficients in $\boldsymbol{\beta}_{\text{miss}}$. Aside from these two changes, the other components of the true SPM model inherited the same form as the model fitted to the HILDA survey in Section 5, with the true parameter values taken as the posterior median estimate. In particular: 1) the same vector of two random effect terms (intercept and random slope for age) was used in $\boldsymbol{z}_{ij,\text{earn}}$, $\boldsymbol{z}_{ij,\text{nonzero}}$, and $\boldsymbol{z}_{ij,\text{miss}}$; 2) the non-zero income distribution $\mathcal{F}_{+}(\cdot;\mu,\phi)$ was set to the lognormal distribution; 3) a quadratic form was used for $s_{\text{miss}}(\cdot)$;

4) a linear form was used for $s_{\text{round}}(\cdot)$. We provide more details and a table of all the true parameter values in Supplementary Material A.

For the simulation study, we considered a subsample of individuals from the HILDA survey. This was obtained by randomly selecting $n = 200$ individuals from the HILDA survey who had $n_i \geq 5$ measurements, and using only the covariates from this subsample. This led to a total of 1626 observations in each simulated dataset. We simulated 400 datasets in total. With the above simulation design and true parameter values, on average each simulated dataset has around 6-8% missing income responses and between 30 to 40% of the observed income values were zero; both of these features were consistent with what was seen in the actual HILDA dataset.

For each simulated dataset, we fitted four competing models: 1) the SPM model in Section 3 (SPM model); 2) a model with no zero rounding component (No rounding); 3) a SPM model with $s_{\text{miss}}(\cdot)$ omitted from the missingness component (No smooth); 4) a model where the missing data was not modelled at all (No missing). Under this model and with Bayesian MCMC sampling using JAGS, the missing data income responses are thus assumed to be missing at random and multiple imputation performed based on the relevant predictive distribution. This contrasts with Model 3, which also assumes the income responses are missing at random but forms an explicit regression model for the probability of missingness as a function of the observed covariates. We used the same set of prior distributions and sampling scheme for all four models; see Section 3. Note some models do not involve parameters for some components.

We considered four measures to assess performance of the competing models: 1) relative bias, defined as $400^{-1}\sum_{t=1}^{400}|\Psi_0|^{-1}(\hat{\Psi}^t - \Psi_0)$ where $\hat{\Psi}^b$ generically denotes the posterior median estimate for some parameter Ψ in simulated dataset t, and Ψ_0 denotes the corresponding true value; 2) relative mean squared error (MSE), defined as $400^{-1}\sum_{t=1}^{400}\Psi_0^{-2}(\hat{\Psi}^t - \Psi_0)^2$; 3) coverage probability, defined as the number of simulated datasets out of 400 where the 95% highest posterior density (HPD) intervals contained the true parameter; 4) the mean interval score (Gneiting & Raftery 2007). Given a $100(1-\alpha)\%$ HPD interval for a parameter Ψ, denoted as (l_Ψ, u_Ψ), the interval score is calculated as $(u_\Psi - l_\Psi) + 2\alpha^{-1} \times \mathbf{1}(l_\Psi < \Psi_0 < u_\Psi)$ where $\mathbf{1}(\cdot)$ is the indicator function. A lower mean interval score corresponds to better overall performance of the constructed HPD intervals i.e., the interval often contains the true parameter and its width is comparably small.

4.1. Results

It was not surprising to observe that Model 1, given it is the true model, performed best overall across the four components of the SPM model (Figures 2 and 3): for most

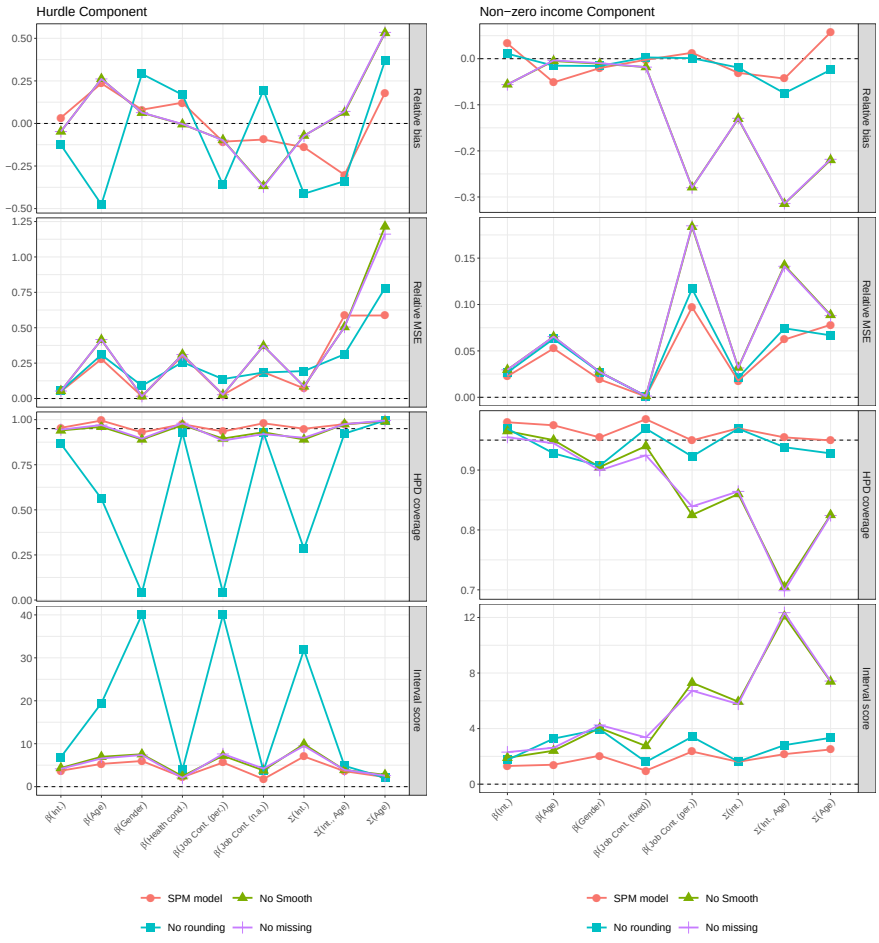
parameters it had the smallest relative bias and relative MSE, the coverage probability of the HPD intervals were always close to the nominal 95% coverage level, and the mean interval score was almost always the lowest of the four models. The most substantial impact of not accounting for zero rounding (Model 2; No rounding) was on inference for the probability of an individual earning an income, where several of the parameters in $\boldsymbol{\Psi}_{\text{earn}}$ incurred comparably high relative bias, coverage probabilities substantially below the nominal level, and the corresponding mean intervals scores being much higher than models that accounted for zero rounding (Figure 2, left column). Inference on the probability of non-response was also affected, but to a lesser extent (Figure 3, left column). Incorrect inference on modelling the probability of an individual earning an income was not surprising given that Model 2 falsely assumes that all observed zeros in the data correspond to truly zero incomes. Turning to Model 3 (No smooth), the failure to model the non-ignorable missingness mechanism impacted inference on the probability of missingness the most (Figures 3, left column). This is expected given that, by not explicitly separating the direct and indirect effects of the covariates, the direct estimates will erroneously incorporate both effects.

Inference on the modelling non-zero income distribution and the probability of zero rounding was also negatively impacted to some extent, and this is to expected to be a consequence of the failure to account for the shared random effects (and thus the correlation) between these component and the missing response component. Finally, given that Model 4 (No missing) also assumes the income is missing at random but does not explicitly model the probability of missingness as a function of the observed covariates, then it was not surprising to see the results from this model were rather similar to those of Model 3. In particular, both inference on the modelling non-zero income distribution and the probability of zero rounding were impacted by this model misspecification.

5. Application to HILDA dataset

We applied the SPM model to a dataset of $n = 1,693$ individuals from waves 1 to 10 (corresponding to 2001 to 2010), from the HILDA survey. The data were relatively unbalanced, with only 481 individuals having records for all 10 waves, while the fewest number of measurements available for an individual was $n_i = 2$ (306 individuals had this cluster size). The average number of measurements for each individual was $n_i = 7.83$, with the number of observations equal to $\sum_{i=1}^{1,693} n_i = 10,650$. As the measure of income and thus the response, we used the answer to the question 'Last financial year, what was your total wage and salary income from all jobs before tax or anything else was deducted?', where the financial year runs from July one calendar year to June the following calendar year. We scaled the responses to be in terms of thousands of Australian dollars. From these observed income

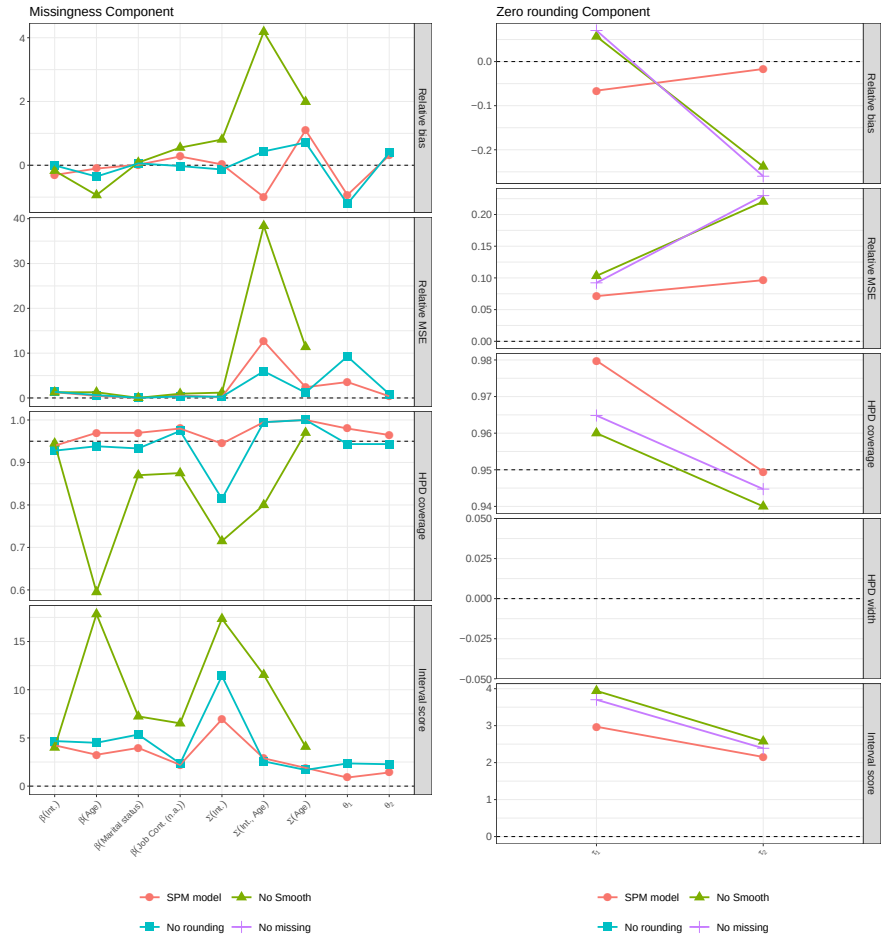
Figure 2. Simulation results for the components of the SPM model relating to the probability of an individual earning an income and the non-zero income distribution (conditional on earning an income). Each row represents one performance measure (from top to bottom: relative bias, relative MSE, coverage probability of 95% HPD intervals, mean interval score). Note some of the four competing models by definition do not contain estimates for certain parameters.



values, there were 809 (7.60%) missing values, 3,459 (32.48%) reported zero incomes, and 6,382 (59.92%) reported non-zero income values.

There was evidence of rounding at several non-zero values of income (Figure 1, right panel), but as discussed at the end of Section 2 the largest repeated non-zero income value only occupied 1.60% of the total number of observations. Of the observed non-zero incomes, the mean and median income (per financial year in thousands of AUD) across all individuals and time points was \$41.468 and \$36.000 respectively, and values ranged from \$0.035 to \$823.241. A histogram of the non-zero income values displayed a strong right skew, while a

Figure 3. Simulation results for the components SPM model relating to the probability of missingness and the probability of zero rounding. Each row represents one performance measure (from top to bottom: relative bias, relative MSE, coverage probability of 95% HPD intervals, mean interval score). Note some of the four competing models by definition do not contain estimates for certain parameters.



log transformation produced a more symmetric distribution close to normality (see Figure 1), although there was some evidence that it was a little too strong with some outliers on the lower end of income values.

We considered a set of five covariates for inclusion in the SPM model: gender (a factor with levels female/male), presence of a long term health condition (a factor with levels no/yes), marital status (a factor with levels single or de facto/divorced/married/separated), job contract (a factor with levels casual/fixed-term/permanent/not asked), and age (a continuous variable ranging from 17 to 93 years, with the median of 41 years and middle 50% of ages occurring between 31 and 54). To clarify, individuals who were categorised as 'not

asked' for job contract type were those who had not undertaken any paid employment in the seven days prior to them answering the survey. But they may still have been employed at some point within the corresponding financial year. There were no missing values for any of the covariates. We provide statistical and graphical summaries of the five covariates in Supplementary Material B.1.

We conducted an exploratory data analysis to decide which covariates to include in the different components of the SPM model. The details of this are provided in Supplementary Material B.2, and we summarise the results final here. As fixed effect covariates in the vectors $\boldsymbol{x}_{ij,\text{earn}}$, $\boldsymbol{x}_{ij,\text{nonzero}}$, and $\boldsymbol{x}_{ij,\text{miss}}$, we included an intercept, all four factor covariates i.e., gender (using female as the reference level), presence of long term health condition (using no as the reference level), marital status (using single or de facto as the reference level), and job contract (using casual as the reference level) as fixed effects using dummy variables, and a smooth term for age defined using thin-plate regression splines. As random effect covariates in the vectors $\boldsymbol{z}_{ij,\text{earn}}$, $\boldsymbol{z}_{ij,\text{nonzero}}$, and $\boldsymbol{z}_{ij,\text{miss}}$, we included a random intercept and slope for age. Note in our exploratory data analysis, the same set of candidate fixed and random effect covariates ended up being included in the components for modelling the probability of earning an income, the distribution of non-zero income, and the probability of non-response.

Finally, one interesting finding from our exploratory analysis was that, ignoring the missing income values, close to half the number of individuals were observed to always have non-zero income (767 out of 1,693; 45.82%), while there were also a sizeable number of individuals who were observed to always have zero income (251 out of 1,693; 14.99%). As a result, we anticipate that the estimated random intercept for the probability of an individual earning an income would be relatively large to accommodate such 'perfect ones' and 'perfect zeros' for the clusters. Analogously, by far the majority of individuals had no missing incomes (1,320 out of 1,693; 78.00%), while there were a small number for which their income was always missing (19 out of 1,693; 1.12%). In turn, we also anticipate that the estimated random intercept for the probability of non-response would also be somewhat large.

5.1. Model fitting and results

We began by fitting the SPM model with three choices for the distribution of non-zero incomes: gamma, exponential and lognormal, as discussed below (2), while fixing $s_{\text{miss}}(\cdot)$ to have a quadratic form; we consider sensitivity analysis of this choice in Supplementary Material B.4. Both the deviance information criterion and penalised expected deviance (Plummer 2008), calculated based only on the observed non-zero incomes, suggested that the lognormal distribution was the most suitable choice for $\mathcal{F}_+(\cdot;\eta,\phi)$, and so we opted to use this distribution for the remainder of the analysis; see the results in Supplementary

Table 1. Posterior median estimates and 95% highest posterior density (HPD) intervals for the parameters in the SPM model fitted to the HILDA survey (excluding the fixed effects of age, which was included as a smooth term), assuming a lognormal distribution for non-zero incomes and a quadratic form for $s_{\text{miss}}(\cdot)$. Fixed effect coefficients whose highest posterior density intervals do not contain zero are marked with an asterisk.

Parameter	Median (HPD interval)	Parameter	Median (HPD interval)
	$\boldsymbol{\Psi}_{\text{earn}}$		$\boldsymbol{\Psi}_{\text{nonzero}}$
Fixed effects ($\boldsymbol{\beta}_{\text{earn}}$)		**Fixed effects** ($\boldsymbol{\beta}_{\text{nonzero}}$)	
Intercept	$5.174\ (4.230, 6.229)^*$	Intercept	$2.670\ (2.545, 2.774)^*$
Gender (male)	$2.091\ (1.643, 2.567)^*$	Gender (male)	$0.736\ (0.648, 0.818)^*$
Health condition (yes)	$-0.536\ (-0.853, -0.241)^*$	Health condition (yes)	$-0.045\ (-0.098, 0.006)$
Marital status (divorced)	$0.434\ (-0.572, 1.442)$	Marital status (divorced)	$0.207\ (-0.051, 0.462)$
Marital status (married)	$0.135\ (-0.336, 0.556)$	Marital status (married)	$-0.011\ (-0.073, 0.046)$
Marital status (separated)	$-0.364\ (-1.115, 0.371)$	Marital status (separated)	$0.050\ (-0.067, 0.172)$
Job contract (fixed-term)	$1.488\ (-0.467, 3.834)$	Job contract (fixed-term)	$0.400\ (0.322, 0.476)^*$
Job contract (permanent)	$1.758\ (0.545, 2.983)^*$	Job contract (permanent)	$0.402\ (0.348, 0.456)^*$
Job contract (not asked)	$-6.862\ (-7.941, -5.968)^*$	Job contract (not asked)	$0.022\ (-0.044, 0.082)$
Random effects ($\boldsymbol{\Sigma}_{\text{earn}}$)		**Random effects** ($\boldsymbol{\Sigma}_{\text{nonzero}}$)	
Intercept	$5.380\ (4.201, 6.770)$	Intercept	$0.624\ (0.552, 0.702)$
Age	$0.522\ (0.145, 1.301)$	Age	$0.456\ (0.362, 0.566)$
Covariance(Int., Age)	$0.897\ (0.207, 1.618)$	Covariance(Int., Age)	$0.330\ (0.264, 0.406)$
		Dispersion parameter	
		ϕ	$0.288\ (0.276, 0.299)$
	$\boldsymbol{\Psi}_{\text{miss}}$		$\boldsymbol{\Psi}_{\text{round}}\ (s_{\text{round}}(\cdot))$
Fixed effects ($\boldsymbol{\beta}_{\text{miss}}$)		τ_0	$-1.696\ (-2.062, -1.400)^*$
Intercept	$-4.342\ (-4.992, -3.683)^*$	τ_1	$-0.105\ (-0.128, -0.074)^*$
Gender (male)	$-0.121\ (-0.514, 0.247)$		
Health condition (yes)	$0.025\ (-0.305, 0.352)$		
Marital status (divorced)	$0.191\ (-1.059, 1.309)$		
Marital status (married)	$0.425\ (0.062, 0.827)^*$		
Marital status (separated)	$0.031\ (-0.650, 0.728)$		
Job contract (fixed-term)	$-0.407\ (-0.992, 0.178)$		
Job contract (permanent)	$-0.223\ (-0.607, 0.151)$		
job contract (not asked)	$-0.759\ (-1.178, -0.328)^*$		
Random effects ($\boldsymbol{\Sigma}_{\text{miss}}$)			
Intercept	$2.480\ (1.746, 3.337)$		
Age	$0.356\ (0.083, 0.913)$		
Covariance(Int., Age)	$-0.199\ (-0.734, 0.329)$		
Smoothing function ($s_{\text{miss}}(\cdot)$)			
θ_1	$-0.009871\ (-0.026275, 0.006357)$		
θ_2	$0.000234\ (0.000115, 0.000324)^*$		

Material B.3. However we note that all three candidate distributions produced similar overall conclusions for the different components of the SPM model. A table of the posterior median estimates and corresponding 95% highest posterior density intervals for the SPM model assuming lognormally distributed non-zero incomes and a quadratic form for $s_{\text{miss}}(\cdot)$ is presented in Table 1, while results with the two alternative choices for $\mathcal{F}_+(\eta, \phi)$ are provided in Supplementary Material B.3.

To visualise different parts of the SPM model, we constructed a set of four graphs. The first two were plots of marginal mean income versus age, where marginal income is defined

as

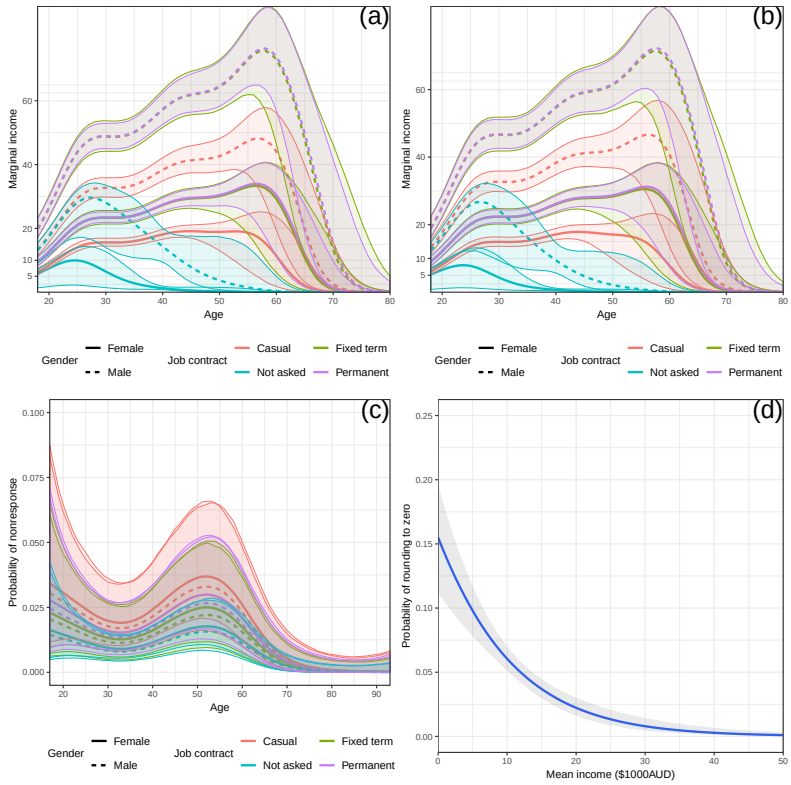
$$\begin{aligned}\mathrm{E}_{Y_{\text{true}}}(y_{ij,\text{true}}|\boldsymbol{b}_{i,\text{earn}}=\mathbf{0},\boldsymbol{b}_{i,\text{nonzero}}=\mathbf{0}) &= \mathrm{E}_G\{\mathrm{E}_{Y_{\text{true}}|G}(y_{ij,\text{true}}|g_{ij},\boldsymbol{b}_{i,\text{earn}}=\mathbf{0},\boldsymbol{b}_{i,\text{nonzero}}=\mathbf{0})\}\\ &= \frac{\exp(\boldsymbol{x}_{ij,\text{earn}}^\top\boldsymbol{\beta}_{\text{earn}})}{1+\exp(\boldsymbol{x}_{ij,\text{earn}}^\top\boldsymbol{\beta}_{\text{earn}})}\exp\left(\boldsymbol{x}_{ij,\text{nonzero}}^\top\boldsymbol{\beta}_{\text{nonzero}}+\frac{1}{2}\phi\right).\end{aligned}$$

Note this is equal to the mean of the mixture model that results from marginalising the non-zero income regression model in (2) with respect to the model for probability of earning an income in (1), and then setting the random effects to zero. Figures 4a,b depict this for individuals with absence and presence of long term health condition, respectively. We also plot the effect of age on the average probability of missing income, $\mathrm{E}(r_{ij}|g_{ij},\boldsymbol{b}_{i,\text{nonzero}}=\mathbf{0},\boldsymbol{b}_{i,\text{miss}}=\mathbf{0})$, in Figure 4c. Finally, a plot of how the probability of rounding to zero ν_{ij} depends on the underlying mean income, μ_{ij}, is provided in Figure 4d.

On average, male individuals who had a permanent job contract (in the past seven days prior to taking the survey) and no long term health condition were substantially more likely to earn an income (Table 1). Individuals who had not undertaken any paid employment in the past seven days prior to taking the survey were substantially less likely to earn income, keeping in mind that income was based on the entire financial year. Conditioning on earning an income, male individuals on a fixed-term or permanent job contract were also on average the highest income earners. There was no clear evidence that marital status had any effect on the propensity to earn income or on the amount of income earned. Finally, both the fixed and random intercept in the model for the probability of earning income were quite large in magnitude: This is in line with what we anticipated from the exploratory data analysis, which found close to half of the individuals observed to have non-zero income, while another 15% of individuals were observed to always report zero income.

From Figures 4a,b, we see that on average marginal income tended to increase in a non-linear manner from age 17, peaking around age 57 before dropping dramatically afterwards to zero. This finding is consistent with previous studies which show income tends to maximise for individuals between the ages 50 to 60. Moreover, the dramatic drop in income after age 60 reflects the probability of earning income itself decreasing rapidly around this age range i.e., retirement age. We caution against over interpreting the model particularly beyond the age of 70, given the data here are comparably sparse; see Supplementary Material B.1. The exception to this trend was individuals who had no job contract (in the past seven days prior undertaking the survey), where marginal income peaks around age 25 before declining to zero. The marginal income plots also clearly showed that for a given age, males tend to have higher incomes than females, while individuals on fixed-term or permanent job contracts have higher incomes than those on casual or no job contracts. Finally, marginal income tended to

Figure 4. Plots illustrating various components of the SPM model fitted to the HILDA survey: (a) marginal income as a function of age, gender, and job contract, for individuals with no long term health condition; (b) marginal income as a function of age, gender, and job contract, for individuals with a long term health condition; (c) probability of missing response as a function of age, gender, and job contract; (d) probability of zero rounding as a function of mean income μ_{ij}. In all plots, 95% pointwise highest posterior density intervals are indicated by shading of the corresponding color.



be slightly lower for individuals with a long term health condition, although this effect was not particularly evident when comparing Figures 4a,b: this was consistent with the fixed effect for health condition being significant (but not large compared to the effects of gender, say) for $\boldsymbol{\Psi}_{\text{earn}}$ but not for $\boldsymbol{\Psi}_{\text{nonzero}}$.

Turning to the model for the probability of missing response, and under the missing data mechanism of the SPM model, we observe a large negative estimated fixed intercept while the random intercept also suggested considerable heterogeneity between individuals (Table 1). This was in line with the exploratory analyses which found over 75% of individuals had no missing incomes at all. Being legally married and not undertaking any paid employment in the seven days prior to undertaking the survey were the only two fixed effect terms found to have substantial direct effects on the probability of non-response. The estimates

of the quadratic smoothing function also provide some evidence that low and high income earners are more likely to not report their income. We explore the sensitivity of the results to the specific missingness assumptions in Supplementary Material B.4. It should be noted that the magnitudes of both estimated coefficients parameterising the quadratic smoothing function were very small. This suggests that the actual effect of the mean response μ_{ij} on the probability of non-response was dominated by the direct effects of covariates. Plotting the overall probability of non-response against age, we see a non-linear effect where probability of missing incomes starts out relatively high at a young age before decreasing and then increasing again, reaching a peak of around 0.05 around ages 53, before decreasing towards zero (Figure 4). This interesting non-linear effect was somewhat consistent with what we saw in our exploratory analysis; see Supplementary Material B.1. More importantly, it should be kept in mind that this non-linear effect is only producing very small changes in the actual probability of missing response e.g., from 0.01 to 0.05. Finally, there was strong evidence of low income earners tended to round and reporting their income as zero (Table 1), with the probability of zero rounding rapidly decreasing towards zero and becoming negligible beyond $20,000 AUD (Figure 4d). Note however the actual probability of zero rounding was at most around 15%.

In Supplementary Material B.4, we present results from a sensitivity analysis used to assess the impact of different choices of $s_{\text{miss}}(\cdot)$, and consequently of different assumptions about the missing data process, on parameter estimates in the SPM model. We adopted an imputation based approach (motivated by Carpenter, Kenward & White 2007, among others) to conduct this assessment. Results show that many of the parameter estimates are indeed sensitive to the assumption of incomes being non-ignorable, with linear and quadratic choices for $s_{\text{miss}}(\cdot)$ tending to produce more similar results to each other compared to assuming $s_{\text{miss}}(\cdot) = 0$. We emphasise that our goal for performing the sensitivity analysis is not to provide evidence in favour of one missingness mechanism over another (since this choice is not explicitly possible to begin with, Molenberghs et al. 2008). Rather, it is simply to show that the HILDA survey is sensitivity to the choice of the missing data assumption.

6. Discussion

We proposed and applied a shared parameter mixture model to analyse national Australian income distributions and evolution over time. The SPM model simultaneously handles the complications of non-ignorable missingness and zero rounding by first modelling the latent, true income process through a two-component mixture model combining a point mass at zero describing the probability of an individual earning an income with a positive continuous distribution describing how non-zero income varies. A shared parameter

component is then used to model the probability of non-response, and where the probability of missingness can vary with the latent non-zero income in a nonlinear manner. Finally, a zero rounding component is included such that the probability of an individual reporting zero income, when they actually earn an income, depends solely on the proximity of the latent income to zero. Applying the SPM model to the HILDA survey produces several key results related to longitudinal income: 1) gender, the presence of a long term health condition, job contract type, and age were important factors in determining the likelihood of earning an income; 2) gender, age, and job contract type have important effects in driving the distribution of non-zero incomes over time; 3) job contract type and marital status had direct effects on the probability of non-response, while the parameter estimate in the SPM model were sensitive to the missingness mechanism assumption; 4) there is a strong association between the probability of zero rounding and a person's true non-zero income.

One of the underlying motivations behind proposing the SPM model was to deal with potential income-dependent non-responses in income. In many datasets on income however, it is quite common to provide imputed values based on assuming the responses are missing at random for practical reasons (as is the case in the HILDA survey e.g., Watson & Starick 2011), even though there is reasonable sociological evidence of missing income responses being non-ignorable (see the references in Section 1 and also Bollinger et al. 2014). While not conclusive, our results from analysing the HILDA survey do show that inference is sensitive to the assumption surrounding the missing data mechanism on income, and so future analyses involving such data from the survey should more seriously acknowledge and/or address this issue. Also, while the use of Bayesian MCMC estimation was strongly motivated by the hierarchical nature of the SPM model, given the increasingly large volume of many longitudinal datasets then future research should explore more efficient estimation approaches (e.g., Kim & Wand 2018; Niku et al. 2019), as well other more efficient Bayesian estimation approaches such as those implemented in Stan (Carpenter et al. 2017) and nimble (de Valpine et al. 2017).

References

BACCI, S. & BARTOLUCCI, F. (2015). A multidimensional finite mixture structural equation model for nonignorable missing responses to test items. *Structural Equation Modeling: A Multidisciplinary Journal* **22**, 352–365.

BACCI, S., BARTOLUCCI, F., BETTIN, G. & PIGINI, C. (2019). A latent class growth model for migrants remittances: an application to the German Socio-Economic Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **182**, 1607–1632.

BECHTEL, L., LORDAN, G. & RAO, D. (2012). Income inequality and mental health – empirical evidence from Australia. *Health economics* **21**, 4–17.

BOHNING, D. & ALFO, M. (2016). Editorial: Special issue on models for continuous data with a spike at zero. *Biometrical Journal* **58**, 255–258.

BOLLINGER, C.R., HIRSCH, B.T., HOKAYEM, C.M. & ZILIAK, J.P. (2014). Trouble in the Tails: Earnings Non-Response and Response Bias across the Distribution. In *Joint Statistical Meetings, Boston, August*.

CALDERN-OJEDA, E., AZPITARTE, F. & GMEZ-DNIZ, E. (2016). Modelling income data using two extensions of the exponential distribution. *Physica A: Statistical Mechanics and its Applications* **461**, 756–766.

CARPENTER, B., GELMAN, A., HOFFMAN, M.D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. & RIDDELL, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software* **76**.

CARPENTER, J.R., KENWARD, M.G. & WHITE, I.R. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research* **16**, 259–275.

CREEMERS, A., HENS, N., AERTS, M., MOLENBERGHS, G., VERBEKE, G. & KENWARD, M.G. (2011). Generalized shared-parameter models and missingness at random. *Statistical modelling* **11**, 279–310.

DE VALPINE, P., TUREK, D., PACIOREK, C.J., ANDERSON-BERGMAN, C., LANG, D.T. & BODIK, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* **26**, 403–413.

DENWOOD, M.J. (2016). runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *Journal of Statistical Software* **71**, 1–25.

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G. & MOLENBERGHS, G. (2008). *Longitudinal data analysis*. Boca Raton, Florida: Chapman and Hall/CRC.

GELMAN, A., JAKULIN, A., PITTAU, M.G. & SU, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2**, 1360–1383.

GELMAN, A. & RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* , 457–472.

GIUSTI, C. & LITTLE, R.J. (2011). An analysis of nonignorable nonresponse to income in a survey with a rotating panel design. *Journal of Official Statistics* **27**, 211–229.

GNEITING, T. & RAFTERY, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

GROSS, M. & RENDTEL, U. (2016). Kernel density estimation for heaped data. *Journal of Survey Statistics and Methodology* **4**, 339–361.

HANISCH, J.U. (2005). Rounded responses to income questions. *Allgemeines Statistisches Archiv* **89**, 39–48.

HEADEY, B. & WOODEN, M. (2004). The effects of wealth and income on subjective well-being and ill-being. *Economic record* **80**, 24–33.

HUI, F.K.C., MLLER, S. & WELSH, A.H. (2017). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica* **27**, 501–518.

JENKINS, S.P. (2010). The British household panel survey and its income data. *ISER Working Paper Series* **33**.

KIM, A.S. & WAND, M.P. (2018). On expectation propagation for generalised, linear and mixed models. *Australian & New Zealand Journal of Statistics* **60**, 75–102.

LITTLE, R.J. & RUBIN, D.B. (2014). *Statistical analysis with missing data*. Hoboken, New Jersey: John Wiley & Sons.

LIU, L., SHIH, Y.C.T., STRAWDERMAN, R.L., ZHANG, D., JOHNSON, B.A. & CHAI, H. (2019). Statistical analysis of zero-inflated nonnegative continuous data: A review. *Statistical Science* **34**, 253–279.

MOLENBERGHS, G., BEUNCKENS, C., SOTTO, C. & KENWARD, M.G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical*

Society: Series B (Statistical Methodology) **70**, 371–388.

NIKU, J., BROOKS, W., HERLIANSYAH, R., HUI, F.K.C., TASKINEN, S. & WARTON, D.I. (2019). Efficient estimation of generalized linear latent variable models. *PloS one* **14**, e0216129.

PLUMMER, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics* **9**, 523–539.

PLUMMER, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*. p. 20–22.

RIPHAHN, R.T. & SERFLING, O. (2005). Item non-response on income and wealth questions. *Empirical Economics* **30**, 521–538.

RIZOPOULOS, D., VERBEKE, G. & MOLENBERGHS, G. (2008). Shared parameter models under random effects misspecification. *Biometrika* **95**, 63–74.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

SCHRAEPLER, J.P. (2006). Explaining income nonresponse – A case study by means of the British Household Panel Study (BHPS). *Quality & Quantity* **40**, 1013–1036.

TSONAKA, R., VERBEKE, G. & LESAFFRE, E. (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics* **65**, 81–87.

WATSON, N. & STARICK, R. (2011). Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey. *Journal of Official Statistics* **27**, 693–715.

WATSON, N. & WOODEN, M.P. (2012). The HILDA survey: a case study in the design and development of a successful household panel survey. *Longitudinal and Life Course Studies* **3**, 369–381.

WULFSOHN, M.S. & TSIATIS, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

ZINN, S. & WRBACH, A. (2016). A statistical approach to address the problem of heaping in self-reported income data. *Journal of Applied Statistics* **43**, 682–703.