Article

JOS

Journal of Official Statistics
2024, Vol. 40(1) 161–189
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0282423X241235252
journals.sagepub.com/home/jof
Sage

Bayesian Inference for Repeated Measures Under Informative Sampling

Terrance D. Savitsky[1],
Luis G. León-Novelo[2], and Helen Engle[3]

Abstract

Survey data are often randomly drawn from an underlying population of inferential interest under a multistage, complex sampling design. A sampling weight proportional to the number of individuals in the population that each sampled individual represents is released. The sampling design is informative with respect to a response variable of interest if the variable correlates with the sampling weights. The distribution for the variables of interest differs in the sample and in the population, requiring correction to the sample distribution to approximate the population. We focus on model-based Bayesian inference for repeated (continuous) measures associated with each sampled individual. We devise a model for the joint estimation of response variable(s) of interest and sampling weights to account for the informative sampling design in a formulation that captures the association of the measures taken on the same individual incorporating individual-specific random-effects. We show that our approach yields correct population inference on the observed sample of units and compare its performance with competing method via simulation. Methods are compared using bias, mean square error, coverage, and length of credible intervals. We demonstrate our approach using a National Health and Nutrition Examination Survey dietary dataset modeling daily protein consumption.

Keywords

mixed effects linear regression, NHANES, pseudolikelihood, sampling weights, survey sampling

[1]Office of Survey Methods Research, Bureau of Labor Statistics, Washington, DC, USA
[2]Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston-School of Public Health, Houston, TX, USA
[3]Thermo Fisher Scientific, Waltham, MA, USA

Corresponding author:
Luis G. León-Novelo, Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston-School of Public Health, 1200 Pressler Street Suite E809, Houston, TX 77030, USA.
Email: Luis.G.LeonNovelo@uth.tmc.edu

1. Introduction

Survey designs for sampling an underlying population of inference often consist of one or more stages to sample clusters of units, followed by the sampling of units. Unequal probabilities of selection are constructed to over-sample some individuals, often to reduce the variance for a domain estimator of interest. A sampled individual who responds to the survey is referred to as a survey participant, or, for simplicity in the sequel, a participant. Inference about the study population needs to consider the sampling design, in particular by incorporating sampling weights into the statistical analysis. Each individual, i, in the population corresponds to a sampling weight, w_i, that is designed to be inversely proportional to the joint inclusion and response probability, π_i, of the individual i as a participant; that is, the individual is selected and responded to the survey. We express this probability mathematically with,

$$\Pr[\text{individual } i \text{ in the population becomes a participant}] \propto \pi_i \propto 1/w_i \tag{1}$$

The weights are, therefore, adjusted for unequal selection probabilities of selection into the survey and for nonresponse, for example, when a selected individual declines to participate. The weights may also be adjusted for other situations; for example, in The National Health and Nutrition Examination Survey (NHANES) dietary datasets released for cycle 2003 to 2004 and later cycles, the dietary sampling weights are adjusted for the day the survey was taken (weekday vs. weekend). We take the perspective of secondary analysts, who are given the weights which are likely to include a nonresponse adjustment by the data producer. In secondary analysis no distinction between sampling and survey response weights is possible and one has to work with the associated unit-level weights. We note that we construct π_i for our modeling in the sequel to be proportional and not necessarily equal to the marginal probability of becoming a participant and, thus, π_i can take any positive value.

Let y_i be the response variable of interest of the individual i in the population. A sampling design is informative with respect to the response variable when the event of becoming a participant and the outcome is related even after conditioning on relevant characteristics of the individual, $\mathbf{v}_i$, which is expressed mathematically by, $y_i \not\perp \pi_i \mid \mathbf{v}_i$. León-Novelo and Savitsky (2019), hereafter referred to as LS2019, propose a model-based Bayesian approach that specifies a joint likelihood for the sampling weights and the response variable of interest to correct for informative sampling. Their approach models the participant probabilities, π_i and the response, y_i, jointly via $p(y_i, \pi_i \mid \boldsymbol{\theta}, \boldsymbol{\kappa}) = p(y_i \mid \boldsymbol{\theta}) p(\pi_i \mid y_i, \boldsymbol{\kappa})$, where $p(y_i \mid \boldsymbol{\theta})$ is the distribution of the response in the study population and $\boldsymbol{\theta}$ is the vector of population parameters of interest, while $\boldsymbol{\kappa}$ the vector of nuisance parameters used to model the relationship between y_i and π_i and serves as an indicator of informativeness for the sampling design (to the extend that the credible interval [CI] for κ_y, an entry of $\boldsymbol{\kappa}$ defined below, is bounded away from 0). The target user for their model formulation to estimate $\boldsymbol{\theta}$ in an unbiased fashion with respect to the population distribution is the data analyst who seeks to estimate the underlying generating parameters $\boldsymbol{\theta}$ from data acquired from a survey sample. It is typical to provide the analyst values of the response variable and predictors for the survey participants along with the associated sampling weights. The approach assumes that the analyst knows the sampling weights and the predictor values for the participants only. The analysts knows neither the sampling weights nor predictor values for non-participants.

In LS2019 the main focus is linear regression with fixed effects. In this article, we extend their approach incorporating random effects in the linear regression model to accommodate repeated measures. Repeated measures arise when a response is measured multiple times for the same participant; for example, the NHANES dietary dataset consists of answers for the same dietary questionnaire at two different days for each participant. Our extension performed in this article incorporates the modeling for the association among the the measures within each participant. This is achieved by constructing participant-specific random effects (P-REs), δ, specified in the marginal linear regression model for the response variable (vs. the conditional model for the sampling weights given the response variable). We consider the case of continuous repeated responses.

The use of random effects to model the correlation among observations is common practice; for example, the NCI method (Tooze et al. 2002, 2006, 2010), which is the approach recommended to estimate typical (daily) nutrient intake when analyzing NHANES dietary data incorporates random effects. In particular, the NCI method is a generalized linear mixed effect model set-up where the correlation of the two repeated measures (i.e., participant nutrient intake in two different days) is modeled by a P-RE. They do not, however, include the sampling weights in their the statistical model. Instead, sampling weights are used to correct for the sampling design when fitting the model via a pseudolikelihood. The contribution of each observation to the log of this pseudolikelihood is proportional to the sampling weight, $\log \text{pseudolikelihood} = \Sigma_i w_i \log p\left(y_i \mid \theta\right)$. Estimation consists of two steps: In the first step, the point estimates maximize the pseudolikelihood. These estimates are asymptotically unbiased. In the second step, confidence intervals for (and/or standard error of) the parameters are calculated via Taylor linearization or re-sampling methods (Centers for Disease Control [CDC] 2016b).

By contrast, our approach incorporates the sampling weights into the likelihood and no second step is required to compute credible intervals for the model parameters in order to achieve correct uncertainty quantification. Ours is the first formulation that incorporates P-REs into the model framework of LS2019.

The NCI method, by contrast, treats the weights as fixed in a “plug-in” formulation, which allows for noise unrelated to the response variable of interest for estimation. The plug-in approach is not fully Bayesian as is our joint modeling formulation such that the uncertainty relative to the distribution over all possible samples is not accounted for. LS2019 show that the pseudo or plug-in likelihood formulation produces overly optimistic or short credible intervals.

A class of pseudolikelihood approaches estimate the parameters of generalized linear mixed models under informative sampling by maximizing the log pseudolikelihood after integrating out the random effects. This approach parameterizes the so-called profile pseudolikelihood. Rabe-Hesketh and Skrondal (2006) propose adaptive quadrature to integrate out the random effects and focus on multistage sampling where random effects are used to model the dependence of units within the same cluster. They mainly focus on logistic regression. Later, Kim et al. (2017) propose an estimation method under informative two-stage cluster sampling. The approach in Kim et al.

(2017) is based on approximating the profile pseudolikelihood using a normal approximation of the sampling distribution of the random effect estimates, avoiding integration of the random effects. Their focus is on linear and logistic regression, while here, ours is on linear regression only.

Their approaches further incorporate repeated measures for the same individual as we propose to do. Their methods use plug-in pseudolikelihood while ours, by contrast, is fully Bayesian using a likelihood defined for the observed sample, rather than approximate pseudolikelihood. Our approach focuses on the estimation of model parameters, $\boldsymbol{\theta}$, of the data generating model and not population totals (e.g., the population average of the response variable). A series of papers Zheng and Little (2003), Little and Zheng (2007), and Zangeneh and Little (2015) propose Bayesian methods to estimate population totals when the inclusion probabilities are proportional to a size variable. All of these approaches estimate the response value of non-sampled units to estimate the population total. By contrast, our approach utilizes only quantities available for sampled units.

In Section 2, we review the basic approach of LS2019. In Section 3 we introduce our extension that incorporates participant-specific random effects. In Section 4, we summarize the pseudolikelihood method and compare its performance with our fully Bayesian formulation in Section 5, in terms of bias, mean square error (MSE), and, coverage, as well as the length of credible intervals. In Section 6, we demonstrate our method with an NHANES dataset, estimating the daily protein consumption in the American population. We conclude with a discussion in Section 7. An Appendix presents details referred to, but not addressed in the main manuscript. We rely on Stan (Carpenter et al. 2016), which performs their No U-turns implementation of the Hamiltonian Monte Carlo posterior sampling algorithm, for estimation of Bayesian hierarchical model posterior distributions estimated in this article.

Going forward, the notation $\text{normal}(\mu,\sigma^2)$ is used to denote the normal distribution with mean μ and variance σ^2 while $\text{normal}(x \mid \mu,\sigma^2)$ denotes its probability density function (PDF) evaluated at x; $\text{lognormal}(\mu,\sigma^2)$ denotes the lognormal distribution, so that $X \sim \text{lognormal}(\mu,\sigma^2)$ is equivalent to $\log X \sim \text{normal}\left(\mu,\sigma^2\right)$ and $\text{lognormal}(x \mid \mu,\sigma^2)$ the respective PDF evaluated at x; $\text{MVN}_p(\mathbf{m},\mathbf{S})$ denotes the p-variate normal distribution with mean vector $\mathbf{m}$ and variance-covariance matrix $\mathbf{S}$; and gamma(a,b) denotes the gamma distribution with shape a and rate b. Matrix, $\boldsymbol{I}_q$, denotes the $q \times q$ identity matrix and $\mathbf{1}_q$ the q dimensional column vector with all *i*ts entries equal to 1. All the non-transposed vectors are column vectors.

2. Review of LS2019 for Single Stage Designs

We next summarize the general formulation of LS2019 that focuses on a single stage of sampling with the model $(\boldsymbol{\theta},\boldsymbol{\kappa})$ parameterized only using fixed effects. We extend and generalize this formation in the next section. Let y_i be the response of the individual i in the population and π_i the corresponding inclusion probability, that is, the probability of s/he becoming a survey participant under the study sampling design (π_i is inversely proportional to the sampling weight w_i). A sampling design is informative for inference on a participant response variable of interest when their inclusion probabilities are correlated with the response variable, $y_i \not\perp \pi_i$ for some i.

LS2019 introduce a Bayesian hierarchical construction that jointly models both the response, y_i, and the marginal inclusion probability, π_i, that is, $p(y_i, \pi_i \mid \boldsymbol{\theta}, \boldsymbol{\kappa}) = p(y_i \mid \boldsymbol{\theta}) \times p(\pi_i \mid y_i, \boldsymbol{\kappa})$, where $p(y_i \mid \boldsymbol{\theta})$ is the response or generating distribution for the population, $\boldsymbol{\theta}$ is the population parameter of interest, and $\boldsymbol{\kappa}$ is the nuisance parameter used to model the relationship between y_i and π_i that provides information on the degree of informativeness of the sample (based on how far the posterior credible intervals are bounded away from 0). LS2019 apply Bayes theorem (see derivation in Appendix A.1 or also Equation (7.1) in Pfeffermann et al. (1998)) to compute

$$p_s(y_i, \pi_i \mid \boldsymbol{\theta}, \boldsymbol{\kappa}) := p\left(\begin{matrix} y_i, \pi_i \mid \boldsymbol{\theta}, \boldsymbol{\kappa}, \text{individual } i \\ \text{is a participant} \end{matrix} \right) = \frac{\pi_i p(\pi_i \mid y_i, \boldsymbol{\kappa})}{E_{y_i^\star \mid \mathrm{q}} \left[E(\pi_i^\star \mid y_i^\star, \boldsymbol{\kappa}) \right]} \times p(y_i \mid \boldsymbol{\theta}). \tag{2}$$

The superindex $\star$ denotes the quantity being integrated out. Note that the denominator in Equation (2) is the marginal probability of individual i becoming a participant. The likelihood for the observed sample,

$$like(\boldsymbol{\theta}, \boldsymbol{\kappa}) = \prod_{i=1}^{n} p_s(y_i, \pi_i \mid \boldsymbol{\theta}, \boldsymbol{\kappa}). \tag{3}$$

We note that for Equation (3) to be a valid likelihood we require

$$\begin{aligned} & p\left[(y_1, \pi_1), \ldots (y_n, \pi_n) \mid \text{individuals 1 to } n \text{ become participants}, \boldsymbol{\theta}, \boldsymbol{\kappa} \right] \\ & = \prod_{i=1}^{n} p\left[(y_i, \pi_i) \mid \text{individual } i \text{ becomes a participant}, \boldsymbol{\theta}, \boldsymbol{\kappa} \right] \end{aligned} \tag{4}$$

Appendix A.2 contains the proof that the following population and design conditions are sufficient for Equation (4):

(C1) $(y_i, \pi_i) \overset{\text{ind}}{\sim} p(\cdot \mid \boldsymbol{\theta}, \boldsymbol{\kappa})$, with index i running over population individuals, are independent. We construct the π_is as unnormalized since a normalization would induce dependence (e.g., if we normalize such that the πs sum to 1, Pr($\pi_2 > 0.5 \mid \pi_1 > 0.6$) = 0 and thus $\pi_1 \not\perp \pi_2$).

(C2) For any individual, conditioned on his/her response and inclusion probability, $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$, the event of becoming a participant (being sampled and responding) is independent of any other individuals becoming participants, their responses and inclusion probabilities.

(C3) Conditioned on $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$, the response and inclusion probability of a population individual is independent of the responses and inclusion probabilities of the participants.

(C3) is natural in our framework, the responses and inclusion probabilities in the population are not affected by the ones in the sample (i.e., the participants). A referee noticed that in practice condition (C1) can be violated if the sampling weights, $w_i \propto 1/\pi_i$, include nonresponse or post-stratification to known population totals adjustment (since the adjustment depends on the common data). For example, if Hispanics tend to have lower response rates, nonresponse adjustment will make their weights higher (and thus dependent), or if the proportion of Whites in the sample is higher than in the population post-stratification adjustment will make their weights lower. If this is the case the analyst is not receiving the π_is as defined in Equation (1) but instead estimates of the π_is that may be dependent. Yet, as secondary analysts we treat these estimates as if they were indeed the independent π_is (despite adjustments for nonresponse and calibration). To cope with this case of dependent (estimated) inclusion probabilities, one can control for the variables used for adjustment (in our example race/ethnicity) when defining the distribution of $\pi_i \mid y_i, \boldsymbol{\kappa}$ such that responses are independent conditioned on $\boldsymbol{\kappa}$ (as we will discuss below after Theorem 1). (C2) is satisfied when sampling is with replacement and non adaptive (i.e., the probability of inclusion does not change by the observed values) but not satisfied when sampling is without replacement from a finite population. Nevertheless, if the population size is much larger than the sample size we can, as it is common practice, approximate the likelihood under sampling without replacement by the likelihood with replacement. When (C1), (C2), and (C3) hold in Equation (3) is a likelihood and the posterior distribution of the model parameters is

$$p_s\left(\boldsymbol{\theta},\boldsymbol{\kappa} \mid \text{data}\right) \propto like\left(\boldsymbol{\theta},\boldsymbol{\kappa}\right) \times Prior\left(\boldsymbol{\theta}\right) \times Prior\left(\boldsymbol{\kappa}\right)$$

where $\text{data} = \left\{(y_i, \pi_i) : i = 1, \ldots, n\right\}$ denotes the sample of size n. Note that without loss of generality, the population individual index i runs from 1 to n in the sample. The formula above allows fully Bayesian inference of the model parameters. The price the modeler pays for this fully Bayesian approach is the requirement to specify a conditional distribution of the inclusion probabilities for all units in the population, $p(\pi_i \mid y_i, \boldsymbol{\kappa})$, and p_s involves complex calculations, namely, the expected value in the denominator of Equation (2). To overcome this, we use STAN and R to estimate the joint posterior distribution for the model parameters. STAN uses Hamiltonian Monte Carlo approach to draw samples from the posterior.

LS2019 jointly model the response and the inclusion probabilities, (y_i, π_i), using only quantities observed in the sample; in particular, the joint distribution of (y_i, π_i) are different in the observed sample and in the population, and we have corrected for this difference in a way that allows us to make unbiased estimation of the parameters of the population model.

Next we review the conditions in LS2019 that produce a mathematically tractable p_s that defines a class of distributions for $p(y_i \mid \boldsymbol{\theta})$ and $p(\pi_i \mid y_i, \boldsymbol{\theta})$ returning a closed form expression for the expectation in the denominator of (2), which simplifies posterior computation. We allow for $p(y_i \mid \boldsymbol{\theta})$ and $p(\pi_i \mid y_i, \boldsymbol{\kappa})$ to depend, respectively, in the set of covariates $\mathbf{u}_i$ and $\mathbf{v}_i$. Since we treat the covariates as fixed (as opposed to random) and to ease notation, we do not explicitly write $p(y_i \mid \boldsymbol{\theta}, \mathbf{u}_i)$ or $p(\pi_i \mid y_i, \boldsymbol{\kappa}, \mathbf{v}_i)$ but instead $p(y_i \mid \boldsymbol{\theta})$ or $p(\pi_i \mid y_i, \boldsymbol{\kappa})$. We also allow that some entries of $\mathbf{u}_i$ overlap $\mathbf{v}_i$, for example both y_i and π_i may depend on gender, or even $\mathbf{u}_i = \mathbf{v}_i$. We now present Theorem 1 in LS2019, that we will adapt to our repeated measurements setting in Theorem 2

Theorem 1. If the population distribution of $\pi_i \mid y_i, \boldsymbol{\kappa}$ is

$$\pi_i \mid y_i, \boldsymbol{\kappa} \sim lognormal\left(\pi_i \mid h\left(y_i, \mathbf{v}_i, \boldsymbol{\kappa}\right), \sigma_\pi^2\right),$$

with the function $h(y_i, \mathbf{v}_i, \boldsymbol{\kappa})$ of the form $h(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) = g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) + t(\mathbf{v}_i, \boldsymbol{\kappa})$ where $\sigma_\pi^2 = \sigma_\pi^2(\boldsymbol{\kappa}, \mathbf{v}_i)$, possibly a function of $(\boldsymbol{\kappa}, \mathbf{v}_i)$ *then*

$$p_s(y_i, \pi_i \mid \boldsymbol{\theta}, \boldsymbol{\kappa}) = \frac{\text{normal}\left(\log \pi_i \mid g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) + t(\mathbf{v}_i, \boldsymbol{\kappa}), \sigma_\pi^2\right)}{\exp\left\{t(\mathbf{v}_i, \boldsymbol{\kappa}) + \sigma_\pi^2 / 2\right\} \times M_y(\boldsymbol{\kappa}; \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta})} \times p(y_i \mid \boldsymbol{\theta})$$

with $M_y(\boldsymbol{\kappa}; \mathbf{u}_i, \mathbf{v}_i, \boldsymbol{\theta}) := E_{y_i^\star \mid \theta}\left[\exp\left\{g(y_i^\star, \mathbf{v}_i, \boldsymbol{\kappa})\right\}\right]$.

Theorem 1 guarantees a closed form expression for p_s in Equation (2) when $p(y_i \mid \boldsymbol{\theta})$ and $M_y(\boldsymbol{\kappa}; \cdots)$ have closed forms. For the particular case of $g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) = \kappa_y y_i$ with $\kappa_y \in \Re$ an entry in $\boldsymbol{\kappa}$, M_y is the moment generating function (MGF) of the population distribution of $y_i \mid \boldsymbol{\theta}$, evaluated at κ_y. Similarly, if we wanted to include an interaction of the response and other covariate, say $v_{i\ell}$, in the model for $\pi_i \mid y_i, \boldsymbol{\kappa}$, we may define $g(y_i, \mathbf{v}_i, \boldsymbol{\kappa}) = \kappa_y y_i + \kappa_{y\ell} v_{i\ell} y_i$ with $\kappa_y, \kappa_{y\ell} \in \Re$ entries of $\boldsymbol{\kappa}$, then M_y is the MGF at evaluated at $\kappa_y + \kappa_{y\ell} v_{i\ell}$. As discussed in LS2019, the assumption of a lognormal distribution for π_i is mathematically attractive since π_i, for individual i, is usually calculated as the product of inclusion probabilities across the stages of the multistage survey design. If each of these stage-wise probabilities are lognormal then their product, $\propto \pi_i$, is lognormal as well.

As long as $t(\mathbf{v}_i, \boldsymbol{\kappa}) = \kappa_0 + \ldots$ contains an intercept term, κ_0, we may assume that π_i is proportional, as opposed to exactly equal, to the inclusion probability for unit i. In other words, no restriction is imposed on $\Sigma_i \pi_i$ where the index i could run over the population or sample indices. This is true since $\pi_i \sim \text{lognormal}(\kappa_0 + \cdots, \cdots)$ implies that $c \times \pi_i \sim \text{lognormal}(\kappa_0 + \log c + \cdots, \cdots)$ where $c > 0$ is any constant and we do not make any inference on the intercept, κ_0. We recommend to include the variables used for nonresponse or post-stratification to population totals adjustments in the vector of covariates $\mathbf{v}_i$ so condition (C1), introduced after Equation (4), holds for the available (estimated) π_is in the sample, that is, $(y_1, \pi_1), \ldots, (y_n, \pi_n) \mid \boldsymbol{\theta}, \boldsymbol{\kappa}$ are independent. $(y_i, \pi_i)s$ are independent if $y_1, \ldots, y_n \mid \boldsymbol{\theta}$ are independent and if $\pi_1, \ldots, \pi_n \mid y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\kappa}$ are independent. The latter independence assumption follows if the relationship between the expected value of log π_i and the adjustment variables is well captured by $h(y_i, \mathbf{v}_i, \boldsymbol{\kappa})$. If adjustments are done, as usually, by multiplying the selection weights, $w_{sel,i} \propto 1/\Pr\left[\text{population individual } i \text{ being invited to participate in the survey}\right]$, by nonresponse and/or post-stratification weights, the linear relationship between log (π_i) and the adjustment variables is appropriate. For example, if nonresponse weights for Hispanics is estimated as the inverse of the response rate for Hispanics RR_H among the individuals invited to participate in the survey, $w_i = w_{sel,i} \times 1/RR_H$, $\log \pi_i = -\log w_i = -\log(w_{sel,i}) + \log(RR_H)$ and $\pi_i \perp \pi_{i'} \mid y_i, y_{i'} \boldsymbol{\theta}, \boldsymbol{\kappa}$ for $i' \neq i$ conditioning on race/ethnicity (included in $\mathbf{v}_i$).

In the sequel, we adapt and extend LS2019 to our particular repeated measurements setup set of conditions in LS2019 on $p(\pi_i \mid y_i, \boldsymbol{\kappa})$ under a likelihood that guarantees the availability of a closed form expression for p_s. This approach assumes that the inclusion probabilities are random, as opposed to the frequentist pseudolikelihood approach discussed later in Section 4 that assumes them fixed.

3. Approach

3.1. Repeated Measures Under Informative Sampling

We consider the mixed effects linear regression population model (for repeated measures),

$$y_{im} = \mathbf{u}_{im}^t \boldsymbol{\beta} + \delta_i + \epsilon_{im} \quad \text{with} \quad \epsilon_{im} \overset{\text{iid}}{\sim} \text{normal}\left(0, \sigma_y^2\right) \text{ and } \delta_i \overset{\text{iid}}{\sim} \text{normal}\left(0, \sigma_\delta^2\right), \tag{5}$$

for each individual i in the population, and $m = 1, \ldots, M_i$, the total number of repeated measures for individual, i. Here, the double index im indexes the population individual i at measurement occasion m; y_{im} is associated value for the response variable; $\mathbf{u}_{im}$ is a q_y dimension vector of covariates whose first entry is set equal to 1 so the model includes an intercept coefficient; and, δ_i is a participant-specific random effect (P-RE).

Denote with $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i,M_i})^t$ the vector of all measurements for individual i and $\mathbf{U}_i = (u_{i1}{}^{q_y \times 1}; \ldots; u_{iM_i})$ the $q_y \times M_i$ matrix whose column m corresponds to covariates at occasion m for individual i. In applications, usually multiple entries of $\mathbf{u}_{im}$ and $\mathbf{u}_{im'}$ naturally match or are exactly equal; for example, when the entry ℓ of $\mathbf{u}_{im}$, $u_{im,\ell}$, encodes the participant's gender or baseline weight, $u_{im,\ell} = u_{im',\ell}$. The population model in Equation (5) is equivalent to

$$\mathbf{y}_i \sim \text{MVN}_{M_i}\left(\mathbf{U}_i^t \boldsymbol{\beta}, \Sigma_{M_i}\right), \quad \text{for individual } i \text{ in the population} \tag{6}$$

with $\Sigma_{M_i} = \sigma_y^2 \boldsymbol{I}_{M_i} + \sigma_\delta^2 \mathbf{1}_{M_i} \mathbf{1}_{M_i}^t$. We parameterize an equal correlation structure but other structures, for example, first order autoregressive, may also be used.

Following LS2019, our Bayesian approach accounts for the informative sampling design by modeling the joint distribution of $(\mathbf{y}_i, \pi_i)$, $p(\mathbf{y}_i, \pi_i \mid \boldsymbol{\theta}, \boldsymbol{\kappa}) = p(\mathbf{y}_i \mid \boldsymbol{\theta}) p(\pi_i \mid \mathbf{y}_i, \boldsymbol{\kappa})$, where $p(\mathbf{y}_i \mid \boldsymbol{\theta})$ is the PDF of the distribution in Equation (6) with $\boldsymbol{\theta} := (\boldsymbol{\beta}, \sigma_\delta, \sigma_y)$; and $p(\pi_i \mid \mathbf{y}_i, \boldsymbol{\kappa})$ is discussed below.

Similar to the set of covariates for $\mathbf{y}_i$, we denote with q_π the number of covariates used to model $\pi_i \mid \mathbf{y}_i, \boldsymbol{\kappa}$; and, $\mathbf{v}_{im}$ the q_π dimensional vector of these covariate values for individual i at occasion m. The first entry of $\mathbf{v}_{im}$ is set equal to 1 to include an intercept and it is common that $v_{im,\ell} = v_{im',\ell}$, where $v_{im,\ell}$ is the entry ℓ of $\mathbf{v}_{im}$. We denote with $\mathbf{V}_i = (\text{v}_{i1}{}^{q_\pi \times 1}; \ldots; \text{v}_{i,M_i})$ the $q_\pi \times M_i$ matrix of covariates. Note that we allow for $\mathbf{v}_{im}$ and $\mathbf{u}_{im}$ to have common covariates or even being equal. For example gender can be used to model both with $\mathbf{y}_i \mid \boldsymbol{\theta}$ and $\pi_i \mid \mathbf{y}_i, \boldsymbol{\kappa}$. Also note that the distributions of $\mathbf{y}_i \mid \boldsymbol{\theta}$ and $\pi_i \mid \mathbf{y}_i, \boldsymbol{\kappa}$ depend, implicitly, on the quantities $\mathbf{u}_i$ and $\mathbf{v}_i$, respectively, but, since they are fixed quantities and to ease notation, we will omit them from the notation of the conditional distributions.

Theorem 2 below presents an extension of Theorem 1 adapted to the repeated measures formulation of Equation (6). The vectors $\mathbf{y}_i$ and $\boldsymbol{\theta}$ in Theorem 2 play the role, respectively, of the univariate response y_i and $\boldsymbol{\theta}$ in Theorem 1. Note that in Theorem 2, we work with the model in Equation (6), where the participant-specific random effects, δ_i, is marginalized to later bring it back in Equation (11).

Theorem 2. If the population distribution of $\pi_i \mid \mathbf{y}_i,\boldsymbol{\kappa}$ is

$$\pi_i \mid \mathbf{y}_i,\boldsymbol{\kappa} \sim lognormal\left(h\left(\mathbf{y}_i,\mathbf{V}_i,\boldsymbol{\kappa}\right),\sigma_\pi^2\right) \tag{7}$$

with the function $h(\mathbf{y}_i,\mathbf{V}_i,\boldsymbol{\kappa})$ of the form $h(\mathbf{y}_i,\mathbf{V}_i,\boldsymbol{\kappa}) = g(\mathbf{y}_i,\mathbf{V}_i,\boldsymbol{\kappa}) + t(\mathbf{V}_i,\boldsymbol{\kappa})$ where $\sigma_\pi^2 = \sigma_\pi^2(\mathbf{V}_i,\mathrm{k})$, possibly a function of $(\mathbf{V}_i,\boldsymbol{\kappa})$ then the exact likelihood for the observed sample takes the form,

$$(\mathbf{y}_i,\pi_i) \mid \theta,\kappa, \text{individual i is a participant} \sim p_s(\mathbf{y}_i,\pi_i \mid \theta,\kappa)$$
$$= \frac{normal\left(\log \pi_i \mid g(\mathbf{y}_i,\mathbf{V}_i,\kappa) + t(\mathbf{V}_i,\kappa),\sigma_\pi^2\right)}{\exp\left\{t(\mathbf{V}_i,\kappa) + \sigma_\pi^2/2\right\} \times M_y(\kappa;\mathbf{U}_i,\mathbf{V}_i,\theta)} \times p(\mathbf{y}_i \mid \theta), \tag{8}$$

with $M_y(\boldsymbol{\kappa};\mathbf{U}_i,\mathbf{V}_i,\boldsymbol{\theta}) := E_{\mathbf{y}_i^\star \mid \theta}\left[\exp\left\{g(\mathbf{y}_i^\star,\mathbf{V}_i,\boldsymbol{\kappa})\right\}\right]$.

Recall that use the superindex $\star$ to denote the quantity being integrated out. We next discuss a common model setting that yields a closed form for Equation (8). If we choose $g(\mathbf{y}_i,\mathbf{V}_i,\boldsymbol{\kappa}) := \kappa_y \bar{y}_{i\cdot}$ with $\bar{y}_{i\cdot} = (1/M_i)\Sigma_{m=1}^{M_i} y_{im}$ the average of the repeated measures of individual i; and, κ_y depending on $\boldsymbol{\kappa}$ and, perhaps, on $\mathbf{V}_i$, then,

$$M_y(\boldsymbol{\kappa};\mathbf{U}_i,\mathbf{V}_i,\boldsymbol{\theta}) := E_{\mathbf{y}_i^\star \mid \theta}\left[\exp\left(\kappa_y \bar{y}_{i\cdot}^\star\right)\right]$$

is the MGF of $\bar{y}_{i\cdot}^\star$ evaluated at κ_y. Under the population model in Equation (6), $\bar{y}_{i\cdot}^\star \sim \text{normal}\left(\bar{\mathbf{u}}_{i\cdot}^t \boldsymbol{\beta}, \sigma_y^2/M_i + \sigma_\delta^2\right)$ with $\bar{\mathbf{u}}_{i\cdot} = \Sigma_{m=1}^{M_i}\mathbf{u}_{im}/M_i$, a column vector of dimension q_y. Since the $\text{normal}(m,s^2)$ distribution has $\text{MGF}(t) = \exp[tm + t^2 s^2/2]$, $M_y(\boldsymbol{\kappa};\mathbf{U}_i,\mathbf{V}_i,\boldsymbol{\theta}) = \exp\left[\kappa_y \bar{\mathbf{u}}_{i\cdot}^t \boldsymbol{\beta} + \kappa_y^2\left(\sigma_y^2/M_i + \sigma_\delta^2\right)/2\right]$, defining $t(\mathbf{V}_i,\mathrm{k}) = \bar{\mathbf{v}}_{i\cdot}^t \mathrm{k}_{\mathbf{v}}$ with $\bar{\mathbf{v}}_{i\cdot} = \Sigma_{m=1}^{M_i}\mathbf{v}_{im}/M_i$, a q_π-dimensional vector, Equation (7) becomes

$$\pi_i \mid \mathbf{y}_i,\boldsymbol{\kappa} \sim \text{lognormal}(\kappa_y \bar{y}_{i\cdot} + \bar{\mathbf{v}}_{i\cdot}^t \boldsymbol{\kappa}_{\mathbf{v}}, \sigma_\pi^2) \tag{9}$$

and Equation (8) becomes

$$p_s(\mathbf{y}_i,\pi_i \mid \boldsymbol{\theta},\boldsymbol{\kappa}) = \left\{\frac{\text{normal}\left(\log \pi_i \mid \kappa_y \bar{y}_{i\cdot} + \bar{\mathbf{v}}_{i\cdot}^t \boldsymbol{\kappa}_{\mathbf{v}}, \sigma_\pi^2\right)}{\exp\left[\bar{\mathbf{v}}_{i\cdot}^t \boldsymbol{\kappa}_{\mathbf{v}} + \sigma_\pi^2/2\right] \times \exp\left[\kappa_y \bar{\mathbf{u}}_{i\cdot}^t \boldsymbol{\beta} + \kappa_y^2\left(\sigma_y^2/M_i + \sigma_\delta^2\right)/2\right]}\right\} \tag{10}$$
$$\times MVN_{M_i}(\mathbf{y}_i \mid \mathbf{U}_i^t \boldsymbol{\beta}, \Sigma_{M_i})$$

with $\boldsymbol{\kappa} = \left(\kappa_y, \boldsymbol{\kappa}_{\mathbf{v}}, \sigma_\pi\right)$ and $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \sigma_y, \sigma_\delta\right)$. Recall that both $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}_{\mathbf{v}}$ include and intercept coefficient. Notice that we could have also used $g\left(\mathbf{y}_i,\mathbf{V}_i,\boldsymbol{\kappa}\right) := \Sigma_m \kappa_{ym} y_{im}$ or $t(\mathbf{V}_i,\boldsymbol{\kappa}) = \Sigma_m \mathbf{v}_{im}^t \boldsymbol{\kappa}_{\mathbf{v}m}$ to give different weights to each repeated measure (response and covariates, respectively) in the distribution of $\log \pi_i \mid \mathbf{y}_i,\boldsymbol{\kappa}$. Similar arguments as the one to derive Equation (10) would give us a close form for p_s in Equation (8).

For ease-of-conducting our simulation study we opt to retain and not marginalize over the participant-specific random effect, treating it as latent variable as an entirely equivalent specification as Equation (10) to obtain,

$$p_s(\mathbf{y}_i,\pi_i \mid \boldsymbol{\theta},\delta_i,\boldsymbol{\kappa}) = \{\cdots\}\prod_{m=1}^{M_i}\text{normal}(y_{im} \mid \mathbf{u}_{im}^t\boldsymbol{\beta}+\delta_i,\sigma_y^2) \text{ with } \delta_i \mid \boldsymbol{\theta} \overset{\text{iid}}{\sim} \text{normal}(0,\sigma_\delta^2) \quad (11)$$

with $\{\cdots\}$ the quantity within curly brackets in Equation (10). The likelihood under $y_i \mid \boldsymbol{\theta},\delta_i$ given in Equation (5) is

$$Like\big(\boldsymbol{\theta},(\delta_1,\ldots,\delta_n),\boldsymbol{\kappa};(\mathbf{y}_1,\pi_1),\ldots,(\mathbf{y}_n,\pi_n)\big) \propto \prod_{i=1}^{n} p_s(\mathbf{y}_i,\pi_i \mid \boldsymbol{\theta},\delta_i,\boldsymbol{\kappa}) \quad (12)$$

The sample size is n and without loss of generality $i=1,\ldots,n$ now indexes the participants (in the observed sample), as opposed to the individual in the population. The expression in Equation (11) represents an augmented likelihood for (y_{im},δ_i) and constructs an augmented posterior distribution when combined with prior distributions for the model global parameters (e.g., $(\boldsymbol{\beta},\boldsymbol{\kappa},\sigma_\pi,\sigma_y)$). The parameter inducing the dependence between $\mathbf{y}_i$ and π_i is κ_y ; and, $\mathbf{y}_i \perp \pi_i \mid \kappa_y = 0$. A 95% credible interval of κ_y non containing zero indicates that the sampling design is informative for the response $\mathbf{y}$. For details on how to define Equation (11) and (12) in Stan code see appendix subsection A.4. In our set-up, since δ_i is latent, we estimate it using the prior distribution $\delta_1,\ldots,\delta_n \mid \boldsymbol{\theta} \overset{\text{iid}}{\sim} \text{normal}(0,\sigma_\delta^2)$ starting with Equation (12). We then proceed to select priors for $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$ to complete the specification of the Bayesian model. We choose the following priors:

$$\begin{aligned}
\delta_1,\ldots,\delta_n \mid \boldsymbol{\theta} &\overset{\text{iid}}{\sim} \text{normal}(0,\sigma_\delta^2)\\
\boldsymbol{\beta} &\sim \text{MVN}_{q_y}\left(\mathbf{0},100\mathbf{I}_{q_y}\right)\\
\left(\kappa_y,\boldsymbol{\kappa}_{\text{v}}^t\right)^t &\sim \text{MVN}_{q_\pi+1}\left(\mathbf{0},100\mathbf{I}_{q_\pi+1}\right)\\
\sigma_\pi &\sim \text{normal}^+(0,1)\\
\sigma_y &\sim \text{normal}^+(0,c_y^2)\\
\sigma_\delta &\sim \text{normal}^+(0,c_\delta^2),
\end{aligned} \quad (13)$$

where the priors on the global parameters are chosen to be vague or weakly informative. Here $\text{normal}^+(\mu,\sigma^2)$ denotes the $\text{normal}(\mu,\sigma^2)$ distribution restricted to the positive real line. When implementing, we standardize the inclusion probabilities so that $\Sigma_{i=1}^n 1/\pi_i = \Sigma_i M_i$, the total number of measurements, matching the standardization of the pseudolikelihood approach below (Section 4). This way the π_is are neither too small nor too large so that the prior distribution for σ_π in Equation (13) is vague. The hyperparameters c_y^2 and c_δ^2 are chosen large enough so the priors are vague. For example, c_y^2 is chosen to be larger than the average over m of the sample variances of $\{y_{im} \mid i=1,\ldots,n\}$ and c_δ^2 is chosen to be larger than the sample covariance of $\{(y_{i1},y_{i2}),i=1,\ldots,n\}$. In the next subsection we extend the proposed method to incorporate primary sampling unit information into the statistical analysis based on León-Novelo and Savitsky (2023). If not of interest this subsection may be skipped.

3.2. Including PSU Information into the Analysis

NHANES data are collected through a complex sampling design. First the U.S. is divided into fifteen strata and two primary sampling units (PSUs) are sampled within each stratum. Strata are defined by the intersection of geography with concentrations of minority populations and a PSU is constructed as a county or a group of geographically contiguous counties. The NHANES data are packaged with variables of interest for each survey participant along with the stratum and PSU identifiers to which s/he belongs to as well as sampling weights. NHANES releases masked stratum and PSU information to protect participant's privacy. Every two-year NHANES-data cycle (CDC 2011) releases information obtained from $H = 15$ strata with $n_h = 2$ PSU per stratum.

León-Novelo and Savitsky (2023) incorporate PSU information into the analysis to account for both possible correlations among the responses of individuals in the same PSU and for informative sampling with respect to PSU (i.e., when the probability of sampling the PSU is not independent of the values of the response variable for nested units). Their approach consists of including a PSU-specific RE (PSU-RE) in both the model for the response and the inclusion probability. They show that the inclusion of these random effects produce correct uncertainty quantification, that is, $1-\alpha$ credible intervals with $1-\alpha$ coverage. León-Novelo and Savitsky (2023) do not consider repeated measures. We now further extend the PSU-REs formulation in León-Novelo and Savitsky (2023) to the repeated measures model in Subsection 3.1. This extension includes a participant-specific RE in the model for the response, and PSU-REs in the models for the response and the inclusion probability.

Let J denote the number of PSUs in the sample where $j = 1,\ldots,J$ denotes the PSU index and n_j denotes the number of observations nested in PSU j. We retain the notation from previous sections replacing the subindex, i with ij, where now the index i runs from $1,\ldots,n_j$; M_{ij} now denotes the number of occasions the response was measured for individual i in PSU j; $\mathbf{y}_{ij}$, π_{ij}, and $\mathbf{U}_{ij} = (\mathbf{u}_{ij1};\ldots;\mathbf{u}_{ijM_{ij}})$ denote, respectively, the M_{ij} dimensional vector of repeated response measures, the inclusion probability of individual ij as a participant, and the $q_y \times M_{ij}$ matrix with m^{th} column, $\mathbf{u}_{ij}$, the vector of covariates at occasion m, for the participant i in the PSU j. The first entry of $\mathbf{u}_{ijm}$, $m = 1,\ldots,M_{ij}$ is set to 1 so the model includes an intercept. Adding the PSU-RE, η_j, to the model in Equation (6) yields,

$$\mathbf{y}_{ij} \mid \boldsymbol{\theta},\eta \sim \text{MVN}_{M_{ij}}\left(\mathbf{U}_{ij}^t \boldsymbol{\beta} + \eta_j \mathbf{1}_{M_{ij}}, \Sigma_{M_{ij}}\right) \quad \text{for } i = 1,\ldots,n_j \text{ and } j = 1,\ldots,J \tag{14}$$

with $\Sigma_{M_{ij}} = \sigma_y^2 \text{I}_{M_{ij}} + \sigma_\delta^2 \mathbf{1}_{M_{ij}} \mathbf{1}_{M_{ij}}^t$; $\boldsymbol{\theta} = \left(\boldsymbol{\beta},\sigma_y,\sigma_\delta\right)$; and $\eta_1,\ldots,\eta_J \overset{\text{iid}}{\sim} \text{normal}\left(0,\sigma_\eta^2\right)$.

Adding the PSU-RE, η_j^π, in the model for π_i defined in Equation (9) yields,

$$\pi_{ij} \mid \mathbf{y}_{ij},\boldsymbol{\kappa},\eta_j^\pi \sim \text{lognormal}\left(\kappa_y \bar{y}_{ij\cdot} + \bar{\mathbf{v}}_{ij\cdot}^t \boldsymbol{\kappa}_{\mathbf{v}} + \eta_j^\pi, \sigma_\pi^2\right), \text{ for } i = 1,\ldots,n_j \text{ and } j = 1,\ldots,J \tag{15}$$

with $\boldsymbol{\kappa} = \left(\kappa_y,\boldsymbol{\kappa}_{\mathbf{v}},\sigma_\pi^2\right)$; $\eta_1^\pi,\ldots,\eta_J^\pi \overset{\text{iid}}{\sim} \text{normal}\left(0,\sigma_{\eta^\pi}^2\right)$; $\bar{y}_{ij\cdot} = (1/M_{ij})\Sigma_{m=1}^{M_{ij}} y_{ijm}$ and $\bar{\mathbf{v}}_{ij\cdot} = (1/M_{ij})\Sigma\mathbf{v}_{ijm}$ and $\mathbf{v}_{ijm}$ the vector of covariate at measurement occasion m used to model $\pi_{ij} \mid \mathbf{y}_{ij}, \boldsymbol{\kappa}, \eta_j^{\pi}$. So, reintroducing the P-RE, δ_{ij}, the analogous to Equation (11) is (this is, just replacing $\bar{\mathbf{v}}_{i\cdot}^{t} \boldsymbol{\kappa}_{\mathrm{v}} \to \bar{\mathbf{v}}_{ij\cdot}^{t} \boldsymbol{\kappa}_{\mathrm{v}} + \eta_j^{\pi}$; and, $\mathbf{u}_{im}^{t} \boldsymbol{\beta} \to \mathbf{u}_{ijm}^{t} \boldsymbol{\beta} + \eta_j$ in Equation (11)).

$$p_s(\mathbf{y}_{ij}, \pi_{ij} \mid \boldsymbol{\theta}, \delta_i, \eta_j, \boldsymbol{\kappa}, \eta_j^{\pi}) = \frac{\operatorname{normal}\left(\log \pi_{ij} \mid \kappa_y \bar{y}_{ij\cdot} + \bar{\mathbf{v}}_{ij\cdot}^{t} \boldsymbol{\kappa}_{\mathrm{v}} + \eta_j^{\pi}, \sigma_{\pi}^2\right)}{\exp\left[\bar{\mathbf{v}}_{ij\cdot}^{t} \boldsymbol{\kappa}_{\mathrm{v}} + \eta_j^{\pi} + \sigma_{\pi}^2 / 2\right] \times \exp\left[\kappa_y \left(\bar{\mathbf{u}}_{ij\cdot}^{t} \boldsymbol{\beta} + \eta_j\right) + \kappa_y^2 \left(\sigma_y^2 / M_{ij} + \sigma_{\delta}^2\right) / 2\right]} \times \prod_{m=1}^{M_{ij}} \operatorname{normal}(y_{ijm} \mid \mathbf{u}_{ijm}^{t} \boldsymbol{\beta} + \delta_{ij} + \eta_j, \sigma_y^2)$$

with now $\delta_{1,1}, \ldots, \delta_{n_1,1}, \delta_{1,2}, \ldots, \delta_{n_2,2}, \ldots, \delta_{1J}, \ldots, \delta_{n_J,J} \overset{\text{iid}}{\sim} N(0, \sigma_{\delta}^2)$. Since the qualities of this approach have been reported in León-Novelo and Savitsky (2023), we will not consider it in our simulation section. This model will be fit in the application section with the priors defined in Equation (13) and $\sigma_{\eta^{\pi}}, \sigma_{\eta} \overset{\text{iid}}{\sim} \operatorname{normal}^{+}(0,1)$.

4. Pseudolikelihood

Savitsky and Williams (2019) (see their Theorem 2) propose an approach to incorporate sampling weights and random effects using a plug-in augmented pseudolikelihood that for the repeated measures set-up of Subsection 3.1 is:

$$\prod_{i=1}^{n} \left\{ \left[\prod_{m=1}^{M_i} p(y_{im} \mid \boldsymbol{\theta})^{w_i} \right] \times p\left(\delta_i \mid \sigma_{\delta}^2\right)^{w_i} \right\} \tag{16}$$

with the sampling weights standardized so $\Sigma_{i=1}^{n} M_i w_i = \Sigma_{i=1}^{n} M_i$. In Subsection A.3 we present the original formula in Theorem 2 of Savitsky and Williams (2019) and derive (16) as a specific case of this formula. The contribution of the observation for a unit, y_{im}, to the pseudolikelihood is its PDF (or what it would contribute to the likelihood) exponentiated to its sampling weight, w_i. The prior distribution for the random effects is also exponentiated by sampling weights, w_i. The sampling weights, w_i are standardized so they sum the number of participant/occasion observations. For example if we have 100 participants with two observations the standardized sampling weights must sum $100 \times 2 = 200$, that is, $\Sigma_{i=1}^{n} w_i = 200$.

The observed data pseudolikelihood for y_{im} together with the pseudo prior for the random effects, δ_i, formulate an augmented data likelihood. The participant-specific random effects are used to account for dependence among the repeated measures. Since $[\operatorname{normal}(x \mid \mu, \sigma^2)]^{w} \propto \operatorname{normal}(x \mid \mu, \sigma^2 / w)$ the pseudolikelihood approach in linear regression is equivalent to the regression model:

$$y_{im} = \mathbf{u}_{im}^{t} \boldsymbol{\beta} + \delta_i + \epsilon_{im} \quad \text{with} \quad \epsilon_{im} \sim \operatorname{normal}\left(0, \sigma_y^2 / w_i\right)$$

with $\delta_i \sim \operatorname{normal}\left(0, \sigma_{\delta}^2 / w_i\right)$, for $m = 1, \ldots, M_i$ and $i = 1, \ldots, n$.

The advantages of the pseudolikelihood approach over the proposed fully Bayesian approach are: (A) It incorporates weights into the power term of the likelihood function so that relatively little modifications are performed to the population model sampler to incorporate the pseudo likelihood; (B) Specification of $\pi_i \mid y_i, \cdots$ for the population is not necessary; (C) There is no expected value

$$\Pr\Big(\text{individual } i \text{ becomes a participant} \mid \boldsymbol{\theta}, \boldsymbol{\kappa}\Big) = E_{y_i^\star \mid \boldsymbol{\theta}}\Big[E(\pi_i^\star \mid y_i^\star, \boldsymbol{\kappa})\Big]$$

(the denominator in Equation 2 to compute as in the fully Bayes method. Note than in (C), the inner and outer expectations may depend in a set of covariates $\mathbf{v}_i$ and $\mathbf{u}_i$, respectively.

The disadvantages of the pseudo posterior approach are: (A) It is not fully Bayesian; (B) The sampling weights are only needed for unbiased estimation to the extent that they are dependent on the response variable of interest. Any variation in the weights not related to the response variable represents noise. The pseudo posterior distribution does not discard variation in weights that is independent of the response variable, so information unrelated to the response introduces noise into the estimation of the pseudo posterior distribution; (C) The weights must be normalized to regulate the amount of estimated posterior uncertainty, which is not required for the fully Bayes approach (except to specify a vague prior for π_i as discussed after equation 13); and, (D) The sampling weights are inversely proportional to the inclusion probabilities. The inclusion probabilities represent a distribution that governs the taking of samples from the population that we call the "sampling design" distribution. The resulting credible intervals of the pseudolikelihood do not account for uncertainty with respect to the sampling design distribution because they treat the inclusion probabilities as fixed.

The pseudolikelihood is used here because it is convenient in that the Bayesian data analyst may use the same model and posterior sampling algorithm as defined for the population and only exponentiates the likelihood contributions by the associated sampling weights. While the pseudoposterior is not our recommended (fully Bayesian) method because it is known that it produces incorrect credible intervals, we include it as a comparison to our fully Bayes procedure because it is the commonly used method in practice due to its ease of implementation.

We implement the pseudolikelihood approach in the sequel as a Bayesian version of the NCI method. The pseudolikelihood approach uses one-step estimation, instead of the two-step estimation algorithm of the NCI approach, which propagates uncertainty in estimation of parameters, but is otherwise equivalent. We show that the fully Bayes approach outperforms the pseudolikelihood approach in terms of bias, MSE, and 95% of CI coverage.

5. Simulation Study

We perform a Monte Carlo simulation study to compare the performance of our fully Bayes method in Subsection 3.1 with the pseudolikelihood approach in Section 4. In each Monte Carlo iteration, we generate a population of size $N_{pop} = 10^5$. The information constructed for each individual in the population is its inclusion probability, two repeated measures, and the value of a predictor at each measurement occasion. Next, we generate an informative sample and a simple random sample. The former is analyzed with our fully Bayes method, the pseudolikelihood approach and the model in Equation

(5) that ignores the informativeness of the sample (labeled POP). The simple random sample is analyzed with the model (5) (label this analysis SRS). The SRS is included to serve as a benchmark for point estimation and uncertainty quantification and is compared to methods estimated on the informative sample taken from the same population. For each population and sample we apply the estimation approaches of our fully Bayes method and associated comparator methods, assessing the bias, MSE, and coverage properties. We focus on inference about β_0 as a global parameter of inferential interest. We will repeat steps 1 to 3 below $M = 1000$ (say) times:

1. Generate a population of $N_{pop} = 10^5$ individuals. For $i = 1, \ldots, N_{pop}$, generate:
   (a) inclusion probabilities $\pi_1, \ldots, \pi_{N_{pop}} \overset{\text{iid}}{\sim} \text{gamma}\left(a_\pi = 4, rate = b_\pi = 2\right)$;
   (b) predictors $u_{i1}, u_{i2} \overset{\text{iid}}{\sim} \text{normal}\left(0,1\right)$;
   (c) individual-specific random effects $\gamma_i \overset{\text{iid}}{\sim} \text{normal}\left(0, \sigma_\gamma^2 = 0.3^2\right)$;
   (d) response $y_{im} \sim \text{normal}\left(\mu_{im}, 0.5^2\right)$ with mean $\mu_{i1} = 1 - 0.5u_{i1} + \gamma_i + \pi_i$ and $\mu_{i2} \neq 1 - 0.5u_{i2} + \gamma_i + \pi_i$. Notice that the covariate values are different, $u_{i1} \neq u_{i2}$, but the effect on the response is the same. We are adding π_i to the mean so $y_i \not\perp \pi_i \mid (u_{i1}, u_{i2})$.
2. Generate a simple random sample (SRS) and an informative sample (IS) (without replacement), each of size $n = 100$. The IS contains $(y_{i1}, y_{i2}, u_{i1}, u_{i2}, \pi_i)$ with probability $\pi_i / \Sigma_{i'=1}^{N_{pop}} \pi_{i'}$, while the SRS uses equal probabilities of value, $1/N_{pop}$. Note that the sum of the π_is in the sample (i.e., $\Sigma_{i=1}^{n} \pi_i^{(s)}$) or in the population (i.e., $\Sigma_{i=1}^{N_{pop}} \pi_i$) is not standardized. Large values of y_{i1} and y_{i2} are more likely to be sampled (large value of π_i is associated with large value of y_{im}). The simulation true parameter value of β_0, the intercept, under our regression model in Equation (5) is $\beta_0^{Ana,TRUE} := E(y_{im} \mid u_{i1} = u_{i2} = 0) = 1 + E(\pi_i) = 1 + 4/2 = 3$.
3. Analyze IS with three methods, and also the SRS. All of them assume the analysis model in Equation (5) with $\boldsymbol{\beta} = (\beta_0, \beta_1)^t$, $\mathbf{u}_{im}^t = (1, u_{im})$, and priors specified in Equation (13) with $c_y^2 = c_\delta^2 = 1$.
   (a) FULL: The proposed Bayesian model in Subsection 3.1; with $\mathbf{v}_{im} = \mathbf{u}_{im}$ so $\bar{\mathbf{v}}_i^t := (1, (u_{i1} + u_{i2})/2)$ in Equation (15).
   (b) PSEUDO: Pseudolikelihood approach as described in Section 4, with $w_i \propto 1/\pi_i$.
   (c) POP: Bayesian model in Equation (5), ignoring the π_is.
4. Analyze the SRS with model (5), label this SRS analysis.
5. Compute and store for each one of the three models plus the analysis of the SRS:
   - Point estimate of β_0 equal to the posterior expected value of β_0.
   - Central 95% CI for β_0.

The above simulation study design will tend to allow non-informative inference (that allows unbiased estimation using the uncorrected population model (i.e., POP) for the slope parameter of interest, β_1, for a sufficiently large sample size due to the conditioning on π_i (in μ_i) used to generate y_i. We could have added an interaction term $\pi_i \times u_{i1}$ in μ_{i1} and $\pi_i \times u_{i2}$ in μ_{i2} to bias the estimation of β_1 under the uncorrected model (POP) and require our fully Bayes likelihood for the observed sample of Equation (12) (i.e., FULL) for correct

Table 1. Simulation Scenarios. Values of $a\pi$ and $b\pi$ and $\beta_0^{\text{Ana,TRUE}} := \mathrm{E}(y_{ij} \mid u_{i1} = u_{i2} = 0)$ in the Simulation True Distribution of $\pi_i \overset{\text{iid}}{\sim} \text{gamma}\left(a\pi, b_\pi\right)$ in Simulation Step 1(a). In *S*4 the Design is Non-Informative, that is, $y_i \perp \pi_i$.

Simulation scenario	$a\pi$	$b\pi$	$\beta_0^{Ana,TRUE}$
*S*1 : low variance	4	2	3
*S*2 : high variance	5	1	6
*S*3 : exponential	1	1	2
*S*4 : non informative	4	2	1

inference. For ease-of-understanding, however, we achieve the same benefit by focusing on the global intercept, β_0, which is informatively estimated under the sample design.

For each method, we end up with M point estimates of β_0 and central 95% credible intervals. Call $\tilde{\beta}_0^m$, $m = 1, \ldots, M$ this point estimate. Similarly call the credible interval (L_0^m, U_0^m). Estimate,

- Bias = $average_{\{m=1,\ldots,M\}} \left\{ \tilde{\beta}_0^m - \beta_0^{Ana,TRUE} \right\}$;
- MSE = $average_m \left\{ \left[\tilde{\beta}_0^m - \beta_0^{Ana,TRUE} \right]^2 \right\}$;
- Coverage = Proportion of times that the central 95% credible intervals contained $\beta_0^{Ana,TRUE}$;
- Expected length of central 95% credible intervals = $average_m \{U_0^m - L_0^m\}$.

We extend the simulation to more scenarios by varying the values of of a_π and b_π in step 1(a). The values are given in Table 1. Simulation scenario $S1$ explores the performance of the methods when variance of the inclusion probabilities $\left(a_\pi / b_\pi^2 = 1\right)$ is low; $S2$ when it is high; $S3$ when the distribution of π_i has mode 0, and thus is very different from the lognormal distribution assumed by Full in Equation (9). $S4$ is different from the three other scenarios; here we set $\mu_{im} = 1 - 0.5 x_{im} + \gamma_i$ (excluding π_i) in simulation step 1(d) so the IS generated in step 2 is, actually, non-informative such that sampling weights are not needed to correct the sample model. $S4$ explores the performance of the methods design to analyze informative samples when the design is actually non-informative, and thus the weights, $\propto 1/\pi_i$ are noise.

Table 2 shows the results. FULL and PSEUDO yield similar point inference quality (i.e., similar bias and MSE) but only FULL yields appropriate uncertainty quantification (i.e., CI reaching nominal coverage). In $S1$-$S3$, FULL and PSEUDO perform similar in terms of bias and MSE, but the PSEUDO central 95% CIs undercover because this approach does not account for the uncertainty induced by the sampling design distribution. The FULL CIs coverage is similar to that for the benchmark SRS at the cost of being wider than SRS CIs. The sampling design can produce estimators that are more or less efficient, depending on the construction for inclusion probabilities. In general, the use of strata makes the sampling design more efficient than SRS, but clustering into PSUs (which is done for convenience of survey administration) is less efficient; meaning, that it produces longer credible intervals. In $S1$-$S3$, POP yields, as expected, biased

Table 2. Bias, MSE, Coverage, and Expected Length of Central 95% Credible Intervals Times 1,000 Under Competing Models.

Method	Bias $\times 10^3$	MSE $\times 10^3$	Coverage $\times 10^3$	Length $\times 10^3$
*S*1: low variance, $a\pi = 4$ and $b\pi = 2$				
FULL	−9	17	967	470
PSEUDO	19	17	908	446
POP	501	267	15	478
SRS	0	12	950	434
*S*2: high variance, $a\pi = 5$ and $b\pi = 1$				
FULL	28	67	948	1,080
PSEUDO	33	69	917	882
POP	986	1,039	29	947
SRS	13	54	936	881
*S*3: Exponential $a\pi = 1$ and $b\pi = 1$				
FULL	−6	32	971	853
PSEUDO	69	53	864	690
POP	1,002	1,028	0	583
SRS	−4	13	931	429
*S*4: non-informative $a\pi = 4$ and $b\pi = 2$				
FULL	1	3	951	205
PSEUDO	0	3	927	201
POP	0	2	942	184
SRS	−2	2	945	184

inference, showing the consequences of not adjusting inference for informative design. These scenarios, but particularly *S3* where the simulation true exponential distribution of π_i has mode at zero (while the density of any lognomal distribution evaluated at zero equals zero), show that FULL is robust against the violation of the lognormal distribution of the π_is that is assumed in Equation (9). *S4* shows that FULL and PSEUDO coverage is appropriate even when the sample is non informative.

As a side note, the data generating model in step 1, generates first the inclusion probability π (in 1.a) and then $\mathbf{y}_i \mid \pi_i$ in (1.d) while our proposed method (FULL) models y_i and $\pi_i \mid \mathbf{y}_i$. This may look counter-intuitive but in both, the data generating model and FULL, we are jointly modeling $(\mathbf{y}_i, \pi_i)$. So, it is not important whether $\mathbf{y}$ is generated conditioned on π or the reverse.

6. Application to NHANES

To demonstrate our method, we model daily protein consumption while controlling for race/ethnicity, age, and gender, using the 2017 to 2018 NHANES nutrition dataset (CDC 2016a). NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. NHANES oversamples subgroups of particular public health interest. During 2015 to 2018 NHANES oversampled certain race/ethnic and age groups (Chen et al. 2020). The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The objective of the dietary interview is to obtain detailed dietary intake, for example, daily protein consumption, from NHANES participants. All selected NHANES participants are required for two 24-hour dietary recall interviews. The first dietary recall interview is collected in person and the second by telephone three to ten days later. The amount of meat, fish, milk, and other dairy foods data consumed (in the past twenty-four hours) for Day 1 and Day 2 are provided and NHANES releases the estimate protein intake of each participant at each day.

NHANES recommends using their sampling weights on the Day 2 when analyzing data of participants completing Day 1 and Day 2 dietary recalls. Day 2 weights are available for the 7,641 participants with Day 1 and Day 2 data. Among other adjustments, Day 2 weights adjust for dietary recall data collection, and for weekdays (Monday through Thursday), and weekend (Fridays though Sundays).

The response variable is $y_{im} = \log(prot_{im} + 1)$ where $prot_{im}$ is the NHANES estimated grams (gr) of protein consumed by the participant i during the past twenty-four hours before his/her Day m dietary interview, $i = 1, \ldots, n$ and $m = 1, 2$. Our covariates are race/ethnicity, age, and gender. Male is the gender reference category. Age is categorized in four brackets as [0–19], [20–39], [40–59], and [60–80] years old with [0–19] as reference group. Race/Ethnicity has five categories Mexican American, other Hispanic, non-Hispanic Black, other races, and non-Hispanic White with the latter as reference group.

We fit the models used in the simulation study (that does not include PSUs): FULL, PSEUDO, POP, and also an extension of our method to adjust for PSU (labeled FULL. PSU) as described in Subsection 3.2. We recall that two-year NHANES cycle data contains thirty PSUs. The vector of covariates in Equation (5) in this application is,

$$\begin{aligned}\mathbf{u}^t = &\Big(1, 1\big(gender = \text{Female}\big), 1\big(Age \in [20,39]\big), 1\big(Age \in [40,59]\big), 1\big(Age \in [60,80]\big),\\ &1\big(Race/Eth = \text{Mexican American}\big), 1\big(Race/Eth = \text{Other Hispanic}\big),\\ &1\big(Race/Eth = \text{Non-Hispanic Black}\big), 1\big(Race/Eth = \text{Other Race}\big)\Big)\end{aligned}$$

Here $1(A)$ denotes the indicator function of the individual in the set A . For FULL and FULL.PSU, the covariate vector to model $\pi_i | y_i$, in Equation (9) and (15) respectively, $\mathbf{v}_i$, is set equal to $\mathbf{v}_i = \mathbf{u}_i$.

FULL and FULL.PSU results indicate that the design is informative for protein consumption. As Table 2 shows, for FULL, the posterior mean (central 95% credible intervals) of κ_y , in Equation (9), is −0.17 (−0.22, −0.13); while for FULL.PSU the mean of κ_y , in Equation (15), is −0.31 (−0.36, −0.26). In both cases, the CI for κ_y does not contain zero.

Figure 1 shows that the methods yield different inference. The figure presents violin plots representing the posterior distribution of the grams of protein consumed by participants in the reference group, that is, of $\exp(\beta_0) - 1$, under all competing models. POP, the model ignoring the sampling weights, underestimates. The estimate under PSEUDO is the highest and its CI does not overlap FULL CI. The difference in point estimates between FULL and PSEUDO probably derives from the use of raw, noisy weights in PSEUDO. These noisy weights contain a sufficiently high variance unrelated to the response variable that at the released sample size there is estimation bias. The CIs under FULL and FULL.PSU overlap but FULL.PSU tends to yield higher standard deviations because FULL.PSU accounts for the possible non-independence of the outcomes of indi-

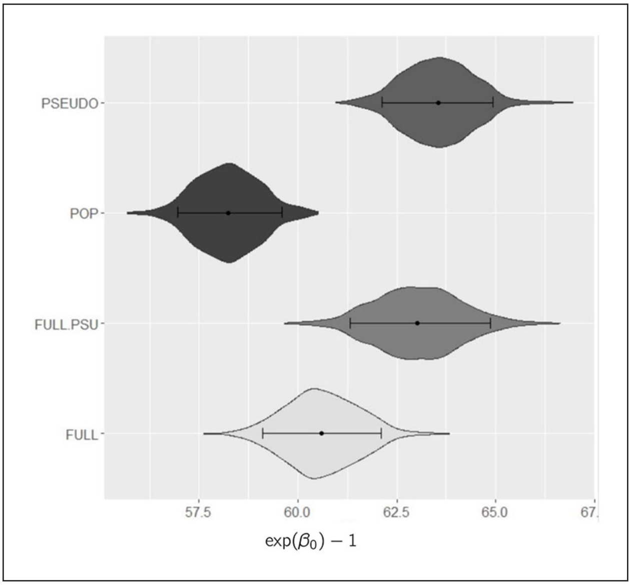

Figure 1. Violin plots along with, mean (dot) and central 95% credible interval (horizontal line) for the expected grams of protein consumed, $\exp(\beta_0)\text{-}1$, under all considered methods in the reference group (White male under 20).

viduals within the same PSU. Since NHANES uses a multi-stage sampling design inference under FULL.PSU is more appropriate.

Table 3 displays mean, standard deviation, and central 95% CI for the model parameters for $y\,|\boldsymbol{\beta},\sigma_y,\sigma_\delta,\ \kappa_y$ under FULL, FULL.PSU and PSEUDO. FULL and FULL.PSU produce similar point estimates except for the intercept for which FULL. PSU yields higher standard error, or equivalently, wider credible interval. This is expected because FULL. PSU takes into account the correlation of the responses within the same PSU (95% CI for σ_η in Equation (14) is $(0.02, 0.05)$). Inference under PSEUDO is different.

7. Final Discussion

LS2019 proposed a method to include the sampling weights into the likelihood to perform Bayesian inference. They mainly focus on the linear regression with fixed effects. We extended their work to account for repeated measures by including a participant-specific random effect and modeling the inclusion probability for individual i , that is, $\pi_i\,|\,\mathbf{y}_i$, as a function the average of all repeated responses for individual i , that is, $\overline{y}_i$, but we could have used any other linear combination of the entries of $\mathbf{y}_i$. Our simulation showed that (A) our proposed method, FULL, yields credible intervals with correct coverage at the cost of wider CI than if we were analyzing a SRS; (B) that this is not always the case for PSEUDO; and (C) that our method is robust against the violation of the lognormal distribution of $\pi_i\,|\,\mathbf{y}_i$ that it assumes. To check (C) the simulation true inclusion probabilities, in Section 5, were generated from gamma and exponential distributions. In LS2019 the robustness against this violation was explored more deeply. For example, in Subsubsection

Table 3. Posterior Mean, Standard Deviation (*SD*), and Central 95% Credible Interval $(q_{0.025}, q_{0.975})$ for Parameters of $y_i \mid \beta, \sigma_y, \sigma_\delta$, and κ_y (to Model $\pi_i \mid \mathbf{y}_i, \kappa_y, \ldots$) Under FULL, FULL.PSU., and PSEUDO.

Parameter	FULL				FULL.PSU				PSEUDO			
	Mean	*SD*	$q_{0.025}$	$q_{0.975}$	Mean	*SD*	$q_{0.025}$	$q_{0.975}$	Mean	*SD*	$q_{0.025}$	$q_{0.975}$
Intercept	4.12	0.02	4.09	4.15	4.16	0.02	4.12	4.19	4.17	0.01	4.14	4.2
Gender female	−0.25	0.01	−0.27	−0.22	−0.25	0.01	−0.27	−0.23	−0.27	0.01	−0.29	−0.25
Age 20–39	0.33	0.02	0.3	0.36	0.33	0.02	0.30	0.36	0.28	0.02	0.25	0.31
Age 40–59	0.31	0.02	0.28	0.34	0.31	0.02	0.28	0.34	0.25	0.02	0.22	0.28
Age 60–80	0.23	0.02	0.2	0.26	0.23	0.02	0.20	0.26	0.21	0.02	0.18	0.24
Mex-American	0.07	0.02	0.03	0.1	0.05	0.02	0.02	0.10	0.04	0.02	0	0.08
Other Hispanic	−0.06	0.02	−0.1	−0.01	−0.07	0.02	−0.11	−0.02	−0.03	0.02	−0.07	0.02
NonHisp Black	−0.08	0.02	−0.11	−0.05	−0.08	0.02	−0.12	−0.07	−0.11	0.02	−0.14	−0.07
Other race	0.02	0.02	−0.01	0.05	0.01	0.02	−0.03	0.05	−0.01	0.02	−0.05	0.03
σ_δ	0.35	0	0.34	0.36	0.35	0	0.34	0.36	0.32	0	0.31	0.33
σ_y	0.44	0	0.43	0.45	0.44	0	0.43	0.45	0.43	0	0.42	0.44
κ_y	−0.17	0.23	−0.22	−0.13	−0.31	0.03	−0.36	−0.26	NNA	NNA	NNA	NNA

4.1.2 they generated the inclusion probabilities, π_is, from a Beta (symmetric) distribution and, in Subsection 4.2, they used the 2013 to 2014 NHANES sampling weights as the simulation truth. Also León-Novelo and Savitsky (2023), in their Subsection 4.2, generated (correlated within cluster) sampling weights from a Dirichlet distribution. In all these simulation scenarios the Fully Bayesian method was robust against the violation of the assumed lognormal distribution of $\pi_i \mid \mathbf{y}_i$. We also incorporated the PSU information following León-Novelo and Savitsky (2023) in Subsection 3.2.

Our method is computationally more expensive than other approaches. We rely on Stan to cope with this limitation. The coding of our method on Stan is simple as shown in the Appendix A.4. The lognormal distribution of $\pi_i \mid \mathbf{y}_i$ with mean a linear combination of the entries of $\mathbf{y}_i$ and other covariates, as shown in Equation (9), remains a computational restriction of our proposed approach. We aim to address this in future work. Another line for future work is to extend the method to binary and count responses.

In this Article, we treat the observed inclusion probabilities as known for the survey participants only. Here the inclusion probability, π_i, is the joint probability of the individual being selected and responding to the survey usually computed assuming independence, as the product of each marginal probability (of inclusion and of responding). The inclusion probabilities, or equivalently the sampling weights, provided to the analyst (e.g., the NHANES dietary publicly available data) are usually adjusted for nonresponse. If the probabilities of being selected were known by the analyst while the one of being a respondent were estimated by the analyst, the estimation error of the latter could be accounted for by modeling the nonresponse probabilities.

A limitation of our simulation study in Section 5 is that it does not incorporate nonresponse explicitly on the data generating process. π_i is defined as the probability of both being selected and responding. We could have generated both the probability of selected to be invited to participate for individual i $\pi_{sel,i}$ and their probability of responding (once invited) as $\pi_{R,i}$. Assuming independence conditioned on the response and relevant covariates, the probability of being selected or invited to participate in the survey and respond is $\pi_i = \pi_{sel,i} \times \pi_{R,i}$, and we would propose to pass only π_i to the data analyst. In our simulation study, we directly generate π_i and pass it to the analyst. Yet, if $\pi_{R,i}$ is estimated from the observed data (e.g., the probability of response for an invited person of Hispanic ethnicity is estimated as the response rate of Hispanics invited to participate), the nonresponse adjustment induces a dependence between the Hispanic participants; nonresponse estimation from common data was not explored in the simulation section. The effect of adjustment for nonresponse when non-response is estimated from common data is a future line of research.

A referee made the observation that if the weights are adjusting for three factors: (1) unequal probability of being invited to participate in the survey, (2) nonresponse, and (3) post-stratification, a more appropriate terminology for these weighs is “survey weights,” while “sampling weights” should be used to refer to weights adjusting for (1) only. We agree, this terminology is more descriptive of what these weights adjust for. We decided to keep the term “sampling weights” to match NHANES terminology where the sampling weights are adjusted for (1), (2), and (3).

In summary, we propose a Bayesian method to the analysis of repeated measures under informative sampling that yields appropriate point estimates (low bias and MSE) and uncertainty quantification (CI reaching nominal coverage).

Acknowledgements

The authors really appreciate the input of the *Journal of Official Statistics* Associate Editor and three referees who carefully reviewed this article and made great suggestions to improve it.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Luis G. Leon-Novelo received financial support for this research from the Houston Health Department/the US Centers for Disease Control and Prevention under award "To Advance Health Equity Through Improved Data and Data Systems for COVID-19 and Other Conditions" [project\#0017758].

References

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. 2016. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 20: 1–37. DOI: https://doi.org/10.18637/jss.v076.i01.

Centers for Disease Control (CDC). 2011. "Continuous NHANES Web Tutorial: Key Concepts About NHANES Survey Design." Available at: https://www.cdc.gov/nchs/tutorials/NHANES/SurveyDesign/SampleDesign/Info1.htm (accessed May, 2018).

Centers for Disease Control (CDC). 2016a. "National Health and Nutrition Examination Survey." Available at: https://www.cdc.gov/nchs/nhanes/index.htm (accessed June 2016).

Centers for Disease Control (CDC). 2016b. "CDC/NCHS/National Health and Nutrition Examination Survey Tutorials. Module 4: Variance Estimation." Available at: https://wwwn.cdc.gov/nchs/nhanes/tutorials/Module4.aspx (accessed January 14, 2021).

Chen, T.-C., J. Clark, M. K. Riddles, L. K. Mohadjer, and T. H. Fakhouri. 2020. "National Health and Nutrition Examination Survey, 2015–2018: Sample Design and Estimation Procedures." DOI: https://pubmed.ncbi.nlm.nih.gov/33663649/.

Kim, J. K., S. Park, and Y. Lee. 2017. "Statistical Inference Using Generalized Linear Mixed Models Under Informative Cluster Sampling." *Canadian Journal of Statistics* 45 (4): 479–97. DOI: https://doi.org/10.1002/cjs.11339.

León-Novelo, L. G., and T. D. Savitsky. 2019. "Fully Bayesian Estimation Under Informative Sampling." *Electronic Journal of Statistics* 13 (1): 1608–45. DOI: https://doi.org/10.1214/19-EJS1538.

León-Novelo, L. G., and T. D. Savitsky. 2023. "Fully Bayesian Estimation Under Dependent and Informative Cluster Sampling." *Journal of Survey Statistics and Methodology* 11 (2): 484–510. DOI: https://doi.org/10.1093/jssam/smab037.

Little, R. J., and H. Zheng. 2007. "The Bayesian Approach to the Analysis of Finite Population Surveys." *Bayesian Statistics* 8 (1): 1–20. DOI: https://doi.org/10.1093/oso/9780199214655.003.0011.

Pfeffermann, D., A. M. Krieger, and Y. Rinott. 1998. "Parametric Distributions of Complex Survey Data Under Informative Probability Sampling." *Statistica Sinica* 8: 1087–114. DOI: http://www.jstor.org/stable/24306526.

Rabe-Hesketh, S., and A. Skrondal. 2006. "Multilevel Modelling of Complex Survey Data." *Journal of the Royal Statistical Society Series A: Statistics in Society* 169 (4): 805–27. DOI: https://doi.org/10.1111/j.1467-985X.2006.00426.x.

Savitsky, T. D., and M. R. Williams. 2019. "Bayesian Mixed Effects Model Estimation Under Informative Sampling." DOI: https://api.semanticscholar.org/CorpusID:119314093.

Tooze, J. A., G. K. Grunwald, and R. H. Jones. 2002. "Analysis of Repeated Measures Data with Clumping at Zero." *Statistical Methods in Medical Research* 11 (4): 341–55. DOI: https://doi.org/10.1191/0962280202sm291ra.

Tooze, J. A., V. Kipnis, D. W. Buckman, R. J. Carroll, L. S. Freedman, P. M. Guenther, S. M. Krebs-Smith, A. F. Subar, and K. W. Dodd. 2010. "A Mixed-Effects Model Approach for

Estimating the Distribution of Usual Intake of Nutrients: The NCI Method." *Statistics in Medicine* 29 (27): 2857–68. DOI: https://doi.org/10.1002/sim.4063.

Tooze, J. A., D. Midthune, K. W. Dodd, L. S. Freedman, S. M. Krebs-Smith, A. F. Subar, P. M. Guenther, R. J. Carroll, and V. Kipnis. 2006. "A New Statistical Method for Estimating the Usual Intake of Episodically Consumed Foods with Application to Their Distribution." *Journal of the American Dietetic Association* 106 (10): 1575–87. DOI: https://doi.org/10.1016/j.jada.2006.07.003.

Zangeneh, S. Z., and R. J. A. Little. 2015. "Bayesian Inference for the Finite Population Total from a Heteroscedastic Probability Proportional to Size Sample." *Journal of Survey Statistics and Methodology* 3 (2): 162–92. DOI: https://doi.org/10.1093/jssam/smv002.

Zheng, H., and R. J. Little. 2003. "Penalized Spline Model-Based Estimation of the Finite Populations Total from Probability-Proportional-to-Size Samples." *Journal of Official Statistics* 19 (2): 99–117. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/penalized-spline-model-based-estimation-of-the-finite-populations-total-from-probability-proportional-to-size-samples.pdf.

Received: May 2022
Accepted: November 2023

A. Appendix

A.1. Derivation of p_s *in Equation (2) Under Informative Sampling*

We construct the distribution of the observed sample taken under an informative design. This approach considers the *population* joint distribution of the response and inclusion probabilities,

$$(y_i,\pi_i)\,|\,\boldsymbol{\theta},\boldsymbol{\kappa} \sim p(y_i,\pi_i\,|\,\boldsymbol{\theta},\boldsymbol{\kappa}) = p(\pi_i\,|\,y_i,\boldsymbol{\kappa})\;p(y_i\,|\,\boldsymbol{\theta}),\quad i=1,\ldots,N, \tag{A1}$$

where N is the population size. Note that we are assuming $\pi_i \perp \boldsymbol{\theta}\,|\,y_i,\boldsymbol{\kappa}$ and $y_i \perp \boldsymbol{\kappa}|\boldsymbol{\theta}$. Let $I_i = 1$ if the individual i in the population becomes a participant and 0 if not. Bayes theorem implies,

$$p(y_i,\pi_i\,|\,\boldsymbol{\theta},\boldsymbol{\kappa},I_i=1) = \frac{\Pr(I_i=1\,|\,y_i,\pi_i,\boldsymbol{\theta},\boldsymbol{\kappa})\times p(y_i,\boldsymbol{\kappa}_i\,|\,\boldsymbol{\theta},\boldsymbol{\kappa})}{\Pr(I_i=1\,|\,\boldsymbol{\theta},\boldsymbol{\kappa})}. \tag{A2}$$

By definition of inclusion probability,

$$\Pr(I_i=1\,|\,y_i,\pi_i,\text{¸},\boldsymbol{\kappa}) = \pi_i \tag{A3}$$

and

$$\begin{aligned}\Pr(I_i=1\,|\,\boldsymbol{\theta},\boldsymbol{\kappa}) &= \iint Pr(I_i=1\,|\,y_i,\pi_i,\boldsymbol{\theta},\boldsymbol{\kappa})p(y_i,\pi_i\,|\,\boldsymbol{\theta},\boldsymbol{\kappa})d\pi_i dy_i\\ &= \iint \pi_i p(\pi_i\,|\,y_i,\boldsymbol{\kappa})d\pi_i\, p(y_i\,|\,\boldsymbol{\theta})dy_i\\ &= \int E(\pi_i\,|\,y_i,\boldsymbol{\kappa})p(y_i\,|\,\boldsymbol{\theta})dy_i\\ &= E_{y_i|\boldsymbol{\theta}}\big[E(\pi_i\,|\,y_i,\boldsymbol{\kappa})\big].\end{aligned} \tag{A4}$$

Plugging (A3) and (A4) into (A2) yields (2)

$$p_s(y_i,\pi_i \mid \boldsymbol{\theta},\boldsymbol{\kappa}) := p(y_i,\pi_i \mid \boldsymbol{\theta},\boldsymbol{\kappa},I_i = 1) = \frac{\pi_i p(\pi_i \mid y_i,\boldsymbol{\kappa})}{E_{y_i^\star \mid \theta}\left[E(\pi_i^\star \mid y_i^\star,\boldsymbol{\kappa})\right]} \times p(y_i \mid \boldsymbol{\theta})$$

Here the superindex $\star$ denotes the quantg integrated out.

A.2. Likelihood in Equation (3) Justification

When we define the likelihood in Equation (3) we are asserting Equation (4) that, introducing the indicator variable $I_i = 1$ when population individual i becomes a participant and $I_i = 0$ otherwise, can be written as

$$p[(y_1,\pi_1),\ldots(y_n,\pi_n) \mid I_1 = I_2 = \ldots = I_n = 1,\boldsymbol{\theta},\boldsymbol{\kappa}] = \prod_{i=1}^{n} p[(y_i,\pi_i) \mid I_i = 1,\boldsymbol{\theta},\boldsymbol{\kappa}] \qquad \text{(A5)}$$

We show below that this independence assertion follows if we assume the following three conditions for the model and the sampling design, respectively, that are the same ones given in Section 2 but in more detail:

(C1) $(y_i,\pi_i) \mid \boldsymbol{\theta},\boldsymbol{\kappa}$, $i = 1,\ldots,N$ are independent in the superpopulation model in (A1).

(C2) For any individual $n+1$, conditioned on his/her response and inclusion probability, $\boldsymbol{\theta}$ and k , the event of becoming the $(n+1)th$ participant is independent of any set of individuals S_n becoming survey participants, their response and inclusion probabilities (if the individual is already in S_n , we mean his/her probability of becoming a participant for a second time under sampling with replacement). Mathematically expressed,

$$\begin{aligned}&\Pr\left[I_{n+1} = 1 \mid (y_{n+1},\pi_{n+1}),\{(y_i,\pi_i) : i \in S_n\},\{I_i = 1 : i \in S_n\},\boldsymbol{\theta},\boldsymbol{\kappa}\right]\\ &= \Pr\left[I_{n+1} = 1 \mid (y_{n+1},\pi_{n+1}),\boldsymbol{\theta},\boldsymbol{\kappa}\right] \propto \pi_{n+1}\end{aligned}$$

where $S_n = \{1,\ldots,n\}$ is the set of indices of first n participants.

(C3) Conditioned on $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$, the response and inclusion probability of a population individual is independent of the response and inclusion probabilities of the n participants already in the survey. In math,

$$\begin{aligned}&p\left[(y_{n+1},\pi_{n+1}) \mid \{(y_i,\pi_i) : i \in S_n\},\{I_i = 1 : i \in S_n\},\boldsymbol{\theta},\boldsymbol{\kappa}\right]\\ &= p\left[(y_{n+1},\pi_{n+1}) \mid \boldsymbol{\theta},\boldsymbol{\kappa}\right]\end{aligned}$$

Proof.

All the probabilities below in this subsection are conditioned on $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$. To ease notation we omit them, that is, we write $p[\cdots \mid \cdots]$ instead of $p[\cdots \mid \cdots,\boldsymbol{\theta},\boldsymbol{\kappa}]$. First we show the following statement that will be helpful in the proof.

$$\Pr\left[I_{n+1} = 1 \mid \{I_i = 1 : i \in S_n\}\right] = Pr\left[I_{n+1} = 1\right] \qquad \text{(A6)}$$

Proof of (A6):

$$
\begin{aligned}
&\Pr\left[I_{n+1}=1\mid\{I_i=1:i\in S_n\}\right]\\
&=\int\Pr\left[I_{n+1}=1\mid(y_{n+1},\pi_{n+1}),\{(y_i,\pi_i):i\in S_n\},\{I_i=1:i\in S_n\}\right]\\
&\qquad\times p\left[(y_{n+1},\pi_{n+1})\mid\{(y_i,\pi_i):i\in S_n\},\{I_i=1:i\in S_n\}\right]\\
&\qquad\times p\left[\{(y_i,\pi_i):i\in S_n\}\mid\{I_i=1:i\in S_n\}\right]d(y_{n+1},\pi_{n+1})d(\{(y_i,\pi_i):i\in S_n\})\\
&=\int\Pr\left[I_{n+1}=1\mid(y_{n+1},\pi_{n+1})\right] &&\text{(Because of (C2))}\\
&\qquad\times p\left[(y_{n+1},\pi_{n+1})\right] &&\text{(Because of (C3))}\\
&\qquad\times p\left[\{(y_i,\pi_i):i\in S_n\}\mid\{I_i=1:i\in S_n\}\right]d(y_{n+1},\pi_{n+1})d(\{(y_i,\pi_i):i\in S_n\})\\
&=\int\Pr\left[I_{n+1}=1\mid(y_{n+1},\pi_{n+1})\right]\times p\left[(y_{n+1},\pi_{n+1})\right]d(y_{n+1},\pi_{n+1})\\
&\qquad\times\int p\left[\{(y_i,\pi_i):i\in S_n\}\mid\{I_i=1:i\in S_n\}\right]d(\{(y_i,\pi_i):i\in S_n\}) &&\text{(This factor equals1)}\\
&=\Pr\left[I_{n+1}=1\right]
\end{aligned}
$$

We prove our assertion by mathematical induction, when the sample size $n=1$ the assertion, that is, (A5), is trivial. We assume the assertion true for sample size n and prove for sample size equal to $n+1$. We need to show

$$
\begin{aligned}
LHS1 &:= p\left[(y_{n+1},\pi_{n+1}),\{(y_i,\pi_i):i\in S_n\}\mid I_{n+1}=1,\{I_i=1:i\in S_n\}\right]\\
&= p\left[(y_{n+1},\pi_{n+1})\mid I_{n+1}=1\right]\times p\left[\{(y_i,\pi_i):i\in S_n\}\mid\{I_i=1:i\in S_n\}\right]=:RHS1
\end{aligned}
$$

Once proven the above statement the assertion is proven since the induction step implies the right factor of $RHS1$ is $\Pi_{i=1}^{n}\ p\left[(y_i,\pi_i)\mid I_i=1\right]$. Applying Bayes Theorem to LHS1 we obtain,

$$
\begin{aligned}
LHS1 &=\Pr\left[I_{n+1}=1\mid(y_{n+1},\pi_{n+1}),\{(y_i,\pi_i):i\in S_n\},\{I_i=1:i\in S_n\}\right]\\
&\qquad\times\underbrace{p\left[(y_{n+1},\pi_{n+1}),\{(y_i,\pi_i):i\in S_n\}\mid\{I_i=1:i\in S_n\}\right]}_{(*)}\\
&\qquad/\Pr\left[I_{n+1}=1\mid\{I_i=1:i\in S_n\}\right]\\
&=\Pr\left[I_{n+1}=1\mid(y_{n+1},\pi_{n+1})\right]\times(*) &&\text{(Because of (C2))}\\
&\qquad/\Pr\left[I_{n+1}=1\right] &&\text{(Because of (A6))}\\
&=\frac{p\left[(y_{n+1},\pi_{n+1})\mid I_{n+1}=1\right]\times\cancel{\Pr\left[I_{n+1}=1\right]}}{p\left[(y_{n+1},\pi_{n+1})\right]}\times\frac{(*)}{\cancel{\Pr\left[I_{n+1}=1\right]}} &&\text{(Bayes Theorem)}\\
&=\frac{p\left[(y_{n+1},\pi_{n+1})\mid I_{n+1}=1\right]}{p\left[(y_{n+1},\pi_{n+1})\right]}\times(*)
\end{aligned}
$$

Applying Bayes Theorem to $(*)$ we obtain

$$(*) = \frac{\overbrace{\Pr\left[\{I_i = 1 : i \in S_n\} \mid (y_{n+1},\pi_{n+1}),\{(y_i,\pi_i) : i \in S_n\}\right]}^{(**)} \times\left\{p\left[(y_{n+1},\pi_{n+1}),\{(y_i,\pi_i) : i \in S_n\}\right]\right\}}{\Pr\left[\{I_i = 1 : i \in S_n\}\right]}$$

$$= (**) \times \frac{\left\{p\left[(y_{n+1},\pi_{n+1})\right] \times p\left[\{(y_i,\pi_i) : i \in S_n\}\right]\right\}}{\Pr\left[\{I_i = 1 : i \in S_n\}\right]} \qquad \text{(Because of (C1))}$$

Applying Bayes Theorem to $(**)$ we obtain

$$(**) = \frac{p\left[(y_{n+1},\pi_{n+1}) \mid \{(y_i,\pi_i) : i \in S_n\},\{I_i = 1 : i \in S_n\}\right]}{p\left[(y_{n+1},\pi_{n+1}) \mid \{(y_i,\pi_i) : i \in S_n\}\right]}$$

$$\times \Pr\left[\{I_i = 1 : i \in S_n\} \mid \{(y_i,\pi_i) : i \in S_n\}\right]$$

$$= \frac{\cancel{p\left[(y_{n+1},\pi_{n+1})\right]} \times Pr\left[\{I_i = 1 : i \in S_n\} \mid \{(y_i,\pi_i) : i \in S_n\}\right]}{\cancel{p\left[(y_{n+1},\pi_{n+1})\right]}} \text{(Because of (C3) and (C1))}$$

$$= \frac{p\left[\{(y_i,\pi_i) : i \in S_n\} \mid \{I_i = 1 : i \in S_n\}\right] \times \Pr\left[\{I_i = 1 : i \in S_n\}\right]}{p\left[\{(y_i,\pi_i) : i \in S_n\}\right]} \qquad \text{(Bayes Theorem)}$$

Replacing the value of $(**)$ above in $(*)$ we obtain

$$(*) = \left(\frac{p\left[\{(y_i,\pi_i) : i \in S_n\} \mid \{I_i = 1 : i \in S_n\}\right] \times \cancel{\Pr\left[\{I_i = 1 : i \in S_n\}\right]}}{\cancel{p\left[\{(y_i,\pi_i) : i \in S_n\}\right]}}\right)$$

$$\times \frac{p\left[(y_{n+1},\pi_{n+1})\right] \times \cancel{p\left[\{(y_i,\pi_i) : i \in S_n\}\right]}}{\cancel{\Pr\left[\{I_i = 1 : i \in S_n\}\right]}}$$

$$= p\left[\{(y_i,\pi_i) : i \in S_n\} \mid \{I_i = 1 : i \in S_n\}\right] \times p\left[(y_{n+1},\pi_{n+1})\right]$$

Replacing the value of $(*)$ above in $LHS1$ we obtain

$$LHS1 = \frac{p\left[(y_{n+1},\pi_{n+1}) \mid I_{n+1} = 1\right]}{\cancel{p\left[(y_{n+1},\pi_{n+1})\right]}} \times \left(\begin{array}{l} p\left[\{(y_i,\pi_i) : i \in S_n\} \mid \{I_i = 1 : i \in S_n\}\right] \\ \times \cancel{p\left[(y_{n+1},\pi_{n+1})\right]} \end{array}\right)$$

$$= p\left[(y_{n+1},\pi_{n+1}) \mid I_{n+1} = 1\right] \times p\left[\{(y_i,\pi_i) : i \in S_n\} \mid \{I_i = 1 : i \in S_n\}\right]$$

$$= RHS1$$

A.3. Pseudolikelihood Approach in Savitsky and Williams

Here we derive the pseudolikelihood approach in Section 4. Savitsky and Williams (2019) considered cluster sampling and in their Theorem 2 they defined the augmented pseudolikelihood as:

$$\prod_{g\in S}\left(\prod_{m\in S_g} p(y_{gm} \mid \theta,\delta_g)^{w_{gm}}\right)p(\delta_g)^{w_g := \frac{1}{|S_g|}\Sigma_{m\in S_g} w_{gm}}$$

with $w_{gm} = 1/\pi_{gm,}$ the specific weight for for unit m nested within cluster g. Here the sample S contains $|S|$ clusters $S_1,\ldots,S_{|S|}$, each cluster with $|S_g|$ observations and δ_g is the cluster specific random effect that models the correlation of the responses y_{gm} within a cluster. The weights are standardized so that $\Sigma_{g\in S} w_{gm} = \Sigma_{g\in S} |S_g|$ the total number of responses. The contribution of the observation for a unit, y_{gm}, to the pseudo-likelihood is its PDF (or what it would contribute to the likelihood) exponentiated to its sampling weight, w_{gm}. The prior distribution for the random effects is exponentiated by cluster-indexed sampling weights, w_g. Each w_g is set equal to the average of the sampling weights of the units nested within the cluster g.

In our context their cluster is a participant. So the index g is i, $|S_g| = M_i$ (the number of observations for the participant) $w_{gm} = w_i \forall m$ and, thus, $w_g = w_i$ and the weights are standardized so that $\Sigma_{i=1}^{n} M_i w_i = \Sigma_{i=1}^{n} M_i$ the total number of participant/occasion measurements. For example if we have 100 participants with 2 observations the standardized sampling weights must sum $100 \times 2 = 200$, that is, $\Sigma_{i=1}^{n} w_i = 200$. Replacing g, S_g, w_{gm} and w_g for i, participant i, w_i and w_i in the equation above yields (16).

A.4. Implementation of (12) Using STAN

We can define the likelihood (12) in Stan in two ways:

1. Directly pass the loglikelihood to the $\log$ of the full joint distribution, in stan, target, in pseudocode,

$$\text{target+} = \log(\mathit{Like}) = \sum_{i=1}^{n} \log p_s(\mathbf{y}_i, \pi_i \mid \cdots)$$

with p_s defined in Equation (11); or,

2. Specify in Stan the distributions of y_{im} and π_i in Equation (5) and (9), respectively, and add to the log of the full joint distribution, referred as target in Stan, the $-\log \Sigma_i \log\big(\text{denominator in Equation } 10\big)$, in pseudocode, this is,

$$\text{target+} = -\left\{\left[\sum_{i=1}^{n} \bar{\mathbf{v}}_i^t\right]\kappa_{\mathrm{v}} + n\sigma_\pi^2/2 + \kappa_y \sum_{i=1}^{n} \bar{\mathbf{u}}_i^t \beta + \kappa_y^2\left[\left(\sum_{i=1}^{n} 1/M_i\right)\sigma_y^2 + n\sigma_\delta^2\right]/2\right\}$$

Stan code, note that the function fortarget_lpd in the code yields the equation above:

```
functions{
   real to_real(int x) { return x;}

   real qxdelta(row_vector x,vector ddelta){
    /*Dot product of x and y*/
      return dot_product(x,ddelta);
   }

   real mupi (real A,real y,row_vector x,vector ddelta) {
    /*mean of pi, conditioned on A,x and delta*/
      return A*y+qxdelta(x,ddelta);
   }

   real muy (row_vector x,vector bbeta,real eta){
    /*mean of y, conditionrd on x,beta and the RE eta*/
      return dot_product(x,bbeta)+eta;
   }

   real fortarget_lpdf (// log of denominator in (9) in paper
       vector pis,     //Vector of inclusion probabilities in the second column
       matrix X_ybar,  //vector of averaged predictors (grouping by participant)
       matrix X_pibar, //Vector of predictors, the first column 1 so the model
                       //for pi includes an intercept
       real suminvj_i, //sum_i 1/number of repeated measures of individual i
       vector bbeta,   //beta: regression coefficients in model for y
       vector ddelta,  //kappa_x: regression coefficient in model for log pi
       real A,         //kappa_y: Coefficient for y in the model for log pi
       vector eta,     //delta: Random effects for model for y
       real sigma2y,   //variance (of the residuals) in the model for y
       real sigma2eta, //variance of random effects
       real sigma2pi   //variance of (of the residuals) in model for log pi
       ){
       int n_individuals=num_elements(pis);
       real sum3=0;
       real sum4=0;
       for(i in 1:n_individuals){
         sum3 += qxdelta(X_pibar[i],ddelta);
         sum4 += dot_product(X_ybar[i],bbeta);
       }
       return(-
                 sum3-
                 n_individuals*sigma2pi/2-
                 A*sum4-
                 A^2*(suminvj_i*sigma2y+n_individuals*sigma2eta)/2
       );
      }
   }//end of the block functions

   data{
       int n_participants;     //Total number of participants
       int nobs;               //length of the response vector
       int p_y;                //number of predictors including the intercept
                               //in the model for y int p_pi;
                               //number of predictors including the
                               //intercept in the model for pi
                               //(inclusion probabilities), besides y
       vector [nobs] ys;       //vector of participant/ocassion measurments
       vector [n_participants] pis;//Vector inclusion probabilities
       int xi [nobs];          //index for REs. xi[i] is the individual to
                               //which the measurement y[i] belongs.
```

```
                            //This is measurement y[j] corresponds
                            //to participant i=xi[j]
    matrix[nobs,p_y] X_y;   //U in paper: matrix of covariates for model for y,
                            //first column must be 1 to include intercept
    matrix[nobs,p_pi] X_pi; //V in paper: Matrix of covariates for model for pi,
                            //first column must be 1 to include intercept
}

transformed data{
    vector [n_participants] ys_bar;
    int    Ms [n_participants]; //Vector of number of observations e.g. Ms[1]=2
                                //indicates individual 1 has 2 repeated measures
    real suminvj_i=0;
    matrix [n_participants,p_y] X_ybar;
    matrix [n_participants,p_pi]X_pibar;

    for(i in 1:n_participants){Ms[i]=0;};
    ys_bar=rep_vector(0.0,n_participants);
    X_ybar = rep_matrix(0.0,n_participants,p_y);
    X_pibar =rep_matrix(0.0,n_participants,p_pi);

    for(j in 1:nobs) {
       ys_bar[xi[j]]+=ys[j];
       Ms[xi[j]]+=+1;
       X_ybar[xi[j]]+=X_y[j];
       X_pibar[xi[j]]+=X_pi[j];
       }

    for(i in 1:n_participants)
       {suminvj_i+=1/to_real(Ms[i]);
       ys_bar[i]=ys_bar[i]/to_real(Ms[i]);
       X_ybar[i]=X_ybar[i]/to_real(Ms[i]);
       X_pibar[i]=X_pibar[i]/to_real(Ms[i]);
       }
}
parameters {
    real <lower=0> sigmapi;  //Standard deviation of residuals in model for log
    real <lower=0> sigmay;   //Standard deviation in residuals in model for y
    real <lower=0> sigma_eta;//Standard deviation of Random effects in model for y

    real A;                  //kappa_y in the paper: regression coef associated
                             //with y in model for log pi Ay+. . .
    vector[p_y] bbeta;       //beta: regression coefficients in model for y
    vector[p_pi] ddelta;    //kappa_x: regression coefficients in model for log pi
    vector[n_participants] eta; //in paper delta_i:
                             //participant-specific RE for model for y
    }

Model{
    bbeta~  normal(0,100);
    ddelta~ normal(0,100);
    A~      normal(0,100);

    eta~ normal(0, sigma_eta);
    sigma_eta normal(0,1);

    sigmapi~ normal(0,1);
    sigmay   normal(0,1);
    for(j in 1:nobs)// Response distributed according to (3) in paper
      ys[j] normal(muy (X_y[j],bbeta,eta[xi[j]]),sigmay);
      }
      for(i in 1:n_participants)//inclusion probabilities distributed as in (8)
      pis[i] lognormal(mupi(A,ys_bar[i],X_pibar[i],ddelta), sigmapi);
      }
      //adding the denominator in (9) to the log of the full joint distribution
      target +=fortarget_lpdf(pis|//Vector of inclusion probabilities
            X_ybar,    //Vector of (averaged) predictors for model for y
            X_pibar,   //Vector of (averaged) predictors for model
                       //for pi ys_bar+X_pibar,
                       //the first column is 1 so the model includes an intercept
            suminvj_i, //sum_i (1/number of repeated measures of individual i)
            bbeta,
            ddelta,A,
            eta,
            sigmay^2,
            sigma_eta^2,
            sigmapi^2);
}
```