A flexible Zero-Augmented Generalized Gamma Mixed Effects Regression Calibration model to Correct for Measurement Error in Episodically Consumed Food

George O. Agogo*

Wageningen University and Research Centre

, Wageningen, Netherlands – george.agogo@wur.nl

Abstract

In nutritional epidemiologic studies, dietary intake is often reported with error in the questionnaire. The dietary intake measurement error usually attenuates the association that relates dietary intake with risk of a disease. As a result, many epidemiologic studies conduct a calibration sub-study to obtain the unbiased measurement for true usual intake. The unbiased measurements can be used to adjust for the error in the questionnaire via regression calibration. Regression calibration predicts the conditional mean of true usual intake given some covariates. Complexities may arise when the unbiased measurements, used as the response in the calibration model, are zero-inflated with skewed-heteroscedastic variance. This is common for foods that are not consumed daily, i.e., episodically consumed foods. We proposed a Zero-augmented Generalized Gamma Mixed Effects Regression Calibration model that can handle the aforementioned complexities and applied the model to NHANES 2003-2004 data. The aim was to investigate the association between fish intake and the blood mercury level. We, further, evaluated the proposed method with the naïve method that ignores error in the reported fish intake. With the naive method, the mercury level is estimated to increase by about 27% per oz intake of fish, whereas with the proposed method, the effect increases by about five fold. In conclusion, the proposed method is able to adjust for the bias caused by measurement error in the questionnaire, when the calibration response cannot be approximated with the standard distributions.

Keywords: bias; covariate measurement error; generalized gamma; zero-inflated models.

1. Introduction

In many epidemiologic studies intake variables are often measured with error. This is rife in nutritional epidemiology, where an individual's usual food intake is measured with error-prone instruments such as food frequency questionnaires (FFQs) (Carroll et al., 2012). Given the long queried period for past intake in the FFQ, people tend to misreport their intake due to memory failure (Willet, 1998). Therefore, using the FFQ measurements to estimate the diet-disease association can lead to substantial bias (Carroll et al., 2012; Kipnis et al., 2009; Tooze et al., 2006).To mitigate this problem, epidemiologic studies often conduct a validation sub-study to take short-term (e.g., for the previous day intake) reference measurement assumed unbiased for true usual intake. The reference measurement is usually reported on a short-term instrument such as 24-hr recall (hereafter, 24HR), mainly administered in a sub-sample of the main-study. Many zeroes are usually recorded in such an instrument for foods that are not consumed daily. The reference measurement can be used to adjust for the bias caused by FFQ measurement error. This can be done with regression calibration (Carroll et al., 2012; Carroll et al., 2006; Fraser and Stram, 2012; Freedman et al., 2008; Kipnis et al., 2009).

Regression calibration involves finding the best conditional mean predictor of true intake, given the FFQ-reported intake and other error-free covariates. The

prediction is used as proxy for true intake to estimate the diet-disease association. Given that true intake is unobservable, the reference intake measurement is used instead in the calibration model. For episodically consumed foods, the commonly faced challenges in regression calibration modelling include how to handle zero-inflation, right-skewness and heteroscedasticity in the reference measurements used as the calibration response. The commonly used approaches are to Box-Cox transform or simply log-transform the non-zero values (Kipnis et al., 2009; Midthune et al., 2011; Tooze et al., 2006) or by simply shifting the response distribution by adding a small constant before transforming (Fraser and Stram, 2012).These approaches usually suffer from complexities in back transforming to the original scale as compounded by the presence of random effects in repeated measures studies(Liu et al., 2010).

In this work, we propose a Zero-augmented Generalized Gamma mixed effects regression calibration model (Manning, Basu and Mullahy, 2005; Yau, Lee and Ng, 2002) to adjust for the bias in the diet-disease association, caused by error in the FFQ. The proposed method is able to handle the aforementioned complexities without transforming the data. The model accounts for heteroscedasticity by permitting the scale parameter to depend on covariate information (Liu et al., 2010; Manning et al., 2005) and provides greater flexibility where simpler models do not permit adequate description of the data. Among the special cases of a generalized gamma distribution include standard gamma, Weibull, exponential and log-normal(Manning et al., 2005). To the best of our knowledge, the proposed model has not been used as a measurement error correction tool. We illustrate the proposed method with the NHANES 2003-2004 data.

2. Regression Calibration to correct for measurement error

Most epidemiologic studies are interested in the association between an intake and risk of a disease. We can assess the association with a GLM model defined by

$$\vartheta\{E(Y_i\,|T_i, \boldsymbol{Z}_i)\} = \beta_T T_i + \boldsymbol{\beta_Z^t Z}_i, \quad (1)$$

where Y_i is disease outcome, T_i is the true dietary intake, $\boldsymbol{Z}_i$ is a vector of error-free covariates for subject i and ϑ is a link function. The interest is in β_T; $\boldsymbol{\beta_Z^t}$ is a vector of coefficient for the error-free covariates. True intake is not observable; therefore, the FFQ measurement is often used leading to an attenuated estimate of β_T. We, further let R_{ij} to denote the j[th] replicate of short-term reference measurement for the i[th] subject from a validation sub-study. The R_{ij} are assumed unbiased for T_i:

$$E\left(R_{ij}\,|i\right) = T_i. \quad (2)$$

We denote the error- contaminated FFQ measurement for each individual in the main study by Q_i.

We can use R_{ij} to adjust for error in the Q_i with regression calibration. Regression calibration involves finding the best conditional mean predictor of T_i given Q_i and other error-free covariates. We denote regression calibration model by $E(T_i|\boldsymbol{Z}_i, \boldsymbol{C}_i)$, where $\boldsymbol{C}_i$ consists of Q_i and other error-free covariates that conditionally predict T_i given $\boldsymbol{Z}_i$ but not predict Y_i. Given that T_i is unobservable, we use R_{ij} in the regression calibration model such that

$$E(T_i|\boldsymbol{Z}_i, \boldsymbol{C}_i\,) = E\left(R_{ij}\,\middle|\boldsymbol{Z}_i, \boldsymbol{C}_i\,\right). \quad (3)$$

Model (3) assumes a non-differential measurement error in Q_i with respect to the disease outcome Y_i, i.e., $f(Y_i|T_i, Q_i, \boldsymbol{Z}_i) = f(Y_i|T_i, \boldsymbol{Z}_i)$, where $f(.)$ is a density function. When dealing with a dietary intake variable not consumed daily, the short-term R_{ij} are often zero-inflated, right-skewed with heteroscedastic variance. Importantly, using such short-term R_{ij} to relate an episodically consumed food with the risk of a disease might not be appropriate, due to the excess zeroes and large within-subject random intake variation.

Given the semi-continuity of R_{ij}, the right hand side of model (3) can be partitioned into two parts, i.e., the mean probability of a non-zero R_{ij} (hereafter, part

I) and the conditional mean of non-zero R_{ij} (hereafter, part II). Part I can be modeled with either a logistic or a probit mixed effect model, whereas part II can be modeled with a plausible family of densities. In this work, we model part II using a flexible generalized gamma mixed effect model.

2.1. A Generalized Gamma distribution for skewed-heteroscedastic non-zero $\mathbf{R_{ij}}$

Following Liu et al. (2010) and Manning et al. (2005), the density of a generalized gamma probability distribution for a non-zero R_{ij} $(R|R>0)$ is defined by three parameters σ, κ and μ as follows

$$f(R|R>0,\kappa,\mu,\sigma)=\frac{\gamma^{\gamma}}{\sigma_{R|R>0}\,\Gamma(\gamma)\sqrt{\gamma}}\exp[u-\gamma\exp(|\gamma|u)], \quad (4)$$

where $\gamma=|\kappa|^{-2}>0$ and $u=sign(\kappa)[\log(R|R>0)-\mu]/\sigma$, σ is a scale parameter that permits heteroscedasticity in the covariate level and Γ is a gamma function. The mean of a generalized gamma distributed non-zero R_{ij} is

$$E(R|R>0,\kappa,\mu,\sigma)=\exp\{\mu+\frac{\sigma\log(\kappa^2)}{\kappa}+\log[\Gamma(\frac{1}{\kappa^2}+\sigma/\kappa)]-\log[\Gamma(1/\kappa^2)]\}$$

and the variance is $var(R|R>0)=\left\{\exp(\mu)\,\kappa^{\frac{2\sigma}{\kappa}}\right\}^2\left\{\frac{\Gamma\left(\frac{1}{\kappa^2}+\frac{2\sigma}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa^2}\right)}-\left[\frac{\Gamma\left(\frac{1}{\kappa^2}+\frac{2\sigma}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa^2}\right)}\right]^{-2}\right\}$

2.2. A Zero-Augmented Generalized Gamma mixed effects Regression Calibration

The two parts can be modelled specified conditionally on a set of covariates $\boldsymbol{X_i}=\{\boldsymbol{Z}_i,\boldsymbol{C}_i\}$, and correlated random effects $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$. We let $G_i=(\boldsymbol{X_i},\boldsymbol{z_i},\boldsymbol{a_i},\boldsymbol{b_i})$, where $\boldsymbol{z_i}$ denotes a design matrix of random effects. The model is specified as follows:

Part I: We model the mean probability of a non-zero R_{ij} using a logistic mixed effects model:

at least one 24 HR, one FFQ, a blood sample for measurement of serum mercury. In the data set, 14.95% (240) reported consumption on day 1 in 24HR and 13.46% (216) reported fish consumption on day 2. Further, 75.14% (1206) reported no fish consumption on either 24HRs and only 3.55% (57) reported fish consumption on both days. In addition, 3.55% (1553) reported fish consumption in the FFQ. Other variables considered include race, education level and age. The ultimate aim was to assess the association between fish intake and blood mercury level.

We use the proposed method to adjust for measurement error in the FFQ and predict T_i conditional on race, age, education and the FFQ. We fit a model that uses the mean of the two 24HR reported fish intake to predict mercury level (hereafter, mean-24HR method), and another model that uses the FFQ intake measurements (hereafter, FFQ method). We use a linear regression model to investigate the association between fish intake and log mercury level. The linear model was adjusted for race, age and level of education.

5. Results

Figure 1 displays the distribution of fish intake as estimated with the mean of 24HR, FFQ and with the proposed method. The histogram for distribution of fish intake, as estimated with the mean of 24HR, suggests a substantial amount of zeroes (left panel). Further shown is a skewed distribution for fish intake as estimated either with the mean of 24HR or FFQ.

Further, we estimated fish intake with the proposed model using NLMIXED procedure in SAS. We use Gauss-Hermite quadrature method with 10 quadrature points but without adaptive centering. Additionally, we use Newton-Raphson optimization technique with 1000 maximum number of iterations. The data showed no indication of heteroscedasticity. The starting values for the model were obtained by fitting the two parts of the model separately using GLIMMIX procedure for part I and NLMIXED procedure for part II. We introduce correlated random intercepts to account for cross-part correlation and use bootstrap method to estimate the correct standard error. The right-skewed distribution for fish intake is further shown by the distribution of intake as calibrated with the proposed method (Figure 1, right panel).

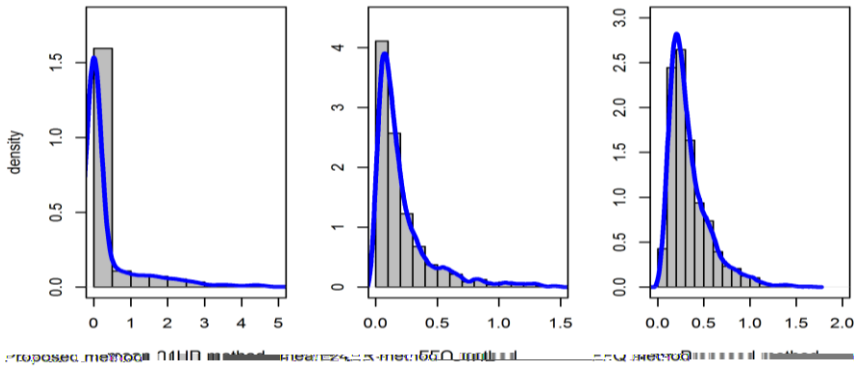

Figure 1. Histograms and density distributions for fish intake as estimated with the mean-24HR, FFQ and the proposed method

Figure 2 displays the least squares line fitted to the scatterplots of the association between fish intake (predicted with the three methods) and log of serum mercury level. Predicting fish intake with the mean of 24HR seriously underestimated the association more than with the FFQ. The proposed method adjusts for the attenuation better than the two methods as shown by steeper slope.

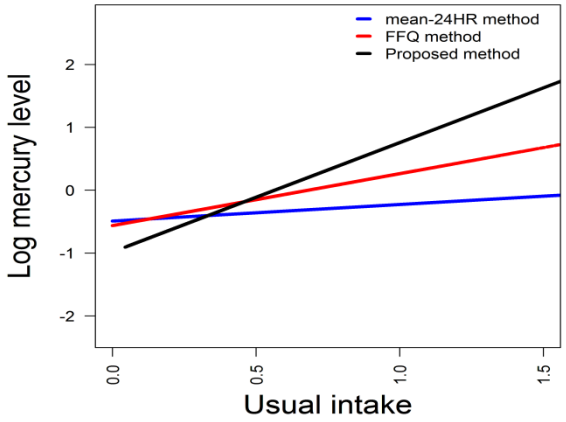

Table 1 presents the estimated coefficients. On average, the estimate obtained with the proposed method is about seven times larger than the one obtained with the mean-24HR and two times larger than the one obtained with the FFQ. The use of the mean of 24HR to estimate fish intake is worse than using FFQ measurements. This is because fish is not consumed daily resulting in many zero 24-HR measurements than in the FFQ given the long queried time for the latter. With the mean-24HR, the mercury level is estimated to increase by about 27% [exp(0.239)] per oz intake of fish, whereas with FFQ the effect increases by about two fold [exp(0.805)]. With the proposed method, the effect increases by about five fold [exp(1.598)]. The proposed method estimates the association with the least precision as expected since it accounts for the uncertainty in the calibration.

Table 1. The regression coefficients estimate for the association between fish intake and log serum mercury level, under the three methods for estimating true usual fish intake, NHANES 2003-2004

Method	$\hat{\beta}_T$	S.E($\hat{\beta}_T$)	R^2
mean-24HR method	0.239	0.0234	0.07
FFQ method	0.805	0.096	0.04
Proposed method	1.533	0.280	0.16

6. Discussion and Conclusion

We proposed a calibration model that is able to the handle zero-inflation, right-skewness and heteroscedasticity in the reference measurements and the cross-part correlation, while adjusting for the bias in the diet-disease association. The ability of the model to adjust for the bias was illustrated with the NHANES data. Importantly, using the mean of short-term reference measurements as an estimate of usual intake for an episodically consumed food to predict disease outcome may be worse than using the FFQ measurements. This is due to large within-person variation and a substantial amount of zeroes evident in the short-term measurements.

The proposed model boasts of its flexibility in modelling data that cannot be handled with simple standard distributions without transforming the data.

Even though we assumed 24 HR measurements as unbiased for true intake, this might not be true as has been shown in previous studies(Kipnis et al., 2003). When the unbiasedness assumption does not hold, the more objective biomarker measurement is preferred. Nevertheless, in the absence of such an objective gold standard measurements, as is the case with most dietary intake, using the 24HR measurements can still provide a better option given the substantial reporting bias usually made in the FFQ. The model suffers from its complexity, compounded by

many parameters being estimated. Therefore, estimation methods such as maximum likelihood can be intractable for complex models posing a serious threat to model convergence. This is the reason why we settled on a random-intercept only model, which in itself might not fit the data adequately. The Bayesian Markov Chain Monte Carlo (MCMC) (Cooper et al., 2007; Zhang et al., 2011), Laplace (Liu et al., 2008) or INLA (Rue, Martino and Chopin, 2009) estimation technique can provide a better alternative. The model, however, needs further evaluation with regard to goodness of fit.

In summary, the proposed method is able to adjust for the bias caused by measurement error in the FFQ, when the calibration response cannot be approximated with the standard distributions.

REFERENCES

Carroll, R. J., Midthune, D., Subar, A. F. (2012). Taking Advantage of the Strengths of 2 Different Dietary Assessment Instruments to Improve Intake Estimates for Nutritional Epidemiology. *&, &□ && **175**, 340-347.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). , *& & & . New York: Chapman& Hall/CRC.

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

Fraser, G. E., and Stram, D. O. (2012). Regression calibration when foods (measured with error) are the variables of interest: markedly non-Gaussian data with many zeroes. *&, &□ && **175**, 325-331.

Freedman, L. S., Midthune, D., Carroll, R. J., and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. **27**, 5195-5216.

Kipnis, V., Midthune, D., Buckman, D. W. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. * **65**, 1003-1010.

Kipnis, V., Subar, A. F., Midthune, D.