Modeling Zero-inflated Clustered Count Data: A Semiparametric Approach

Kevin Carl P. Santos
School of Statistics
University of the Philippines Diliman

This paper proposes to use an additive semiparametric Poisson regression in modeling zero-inflated clustered data. Two estimation methods are exploited in this paper based on de Vera (2010). The first simultaneously estimates both the parametric and nonparametric parts of the model. The second utilizes the backfitting algorithm by smoothing the nonparametric function of the covariates and then estimating the parametric parts of the postulated model. The predictive accuracy, measured in terms of root mean square error (RMSE), of the proposed methods is compared to that of ordinary Zero-Inflated Poisson (ZIP) regression model. Through a simulation study, the average RMSE of the ordinary ZIP regression model is at most 81% and 27% higher for equal and unequal cluster sizes, respectively, than that of proposed model whose parametric and nonparametric parts are simultaneously estimated.

Keywords: *Zero-Inflated Poisson models, clustered data, Generalized Additive Models, backfitting algorithm*

1. Introduction

A wide range of disciplines encounters count data. In genetics, Naya et al. (2008) studied the number of dark spots in Corriedale sheep in Uruguay. Lee et al. (2007) conducted a research on the factors affecting the United States (US) patent citation counts belonging to Korea Institute of Science and Technology (KIST). Jang et al. (2010) analyzed the road safety countermeasures using the data on the number of accidents of the National Highway Number 26 in Korea. Moreover, Baghban et al. (2013) studied the economic costs imposed by gastro-esophageal reflux disease (GERD) and dyspepsia while Barry and Welsh (2002) investigated the abundance of organisms in wildlife populations. The number of claims in a private health insurance scheme is, meanwhile, the interest of Mouatassim and Ezzahid (2012). Hence, modeling these kind of data is certainly necessary.

The most common statistical modeling technique used in analyzing count data is Poisson regression. Poisson regression is a special case of the generalized linear model (GLM) wherein the logarithm link is used to model its expected value. It has equidispersion assumption, i.e., the mean and variance are equal. However, if there are excess zeros, this assumption is easily not satisfied, and, hence, the model will not fit the data well. To solve this, family of models handling zero-inflation emerged such as Zero-Inflated Poisson (ZIP) Regression Model. This model combines two processes – a model for excess zeros and another for nonzero values of the response variable.

In the recent years, ZIP regression model is extended to clustered data as it affects hypothesis testing results and estimation procedures. Lee et al. (2011) proposed a marginalized model approach in analyzing zero-inflated clustered count data using random effects to explain within-subject dependence. They utilized Quasi-Newton algorithm in the estimation procedure. Their proposed method uses conditional models for serial association of responses while still modeling the marginal mean as a function of covariates directly. Thus, the specification of the dependence model does not affect the interpretability of the regression coefficients in two marginal components. However, this method assumes parametric form of the model.

Furthermore, Demidenko (2007) investigated different methods in estimating the parameters of a Poisson regression model intended for clustered data. One of the methods is ordinary Poisson regression with random cluster-specific intercept, i.e., $Y_{ij} \sim P_o(\mu_i + X_{ij}'\beta)$ where Y_{ij} refers to the j^{th} observation in the i^{th} cluster and μ_i is the random cluster-specific intercept. The addition of a random intercept accounts for the within-cluster marginal correlation. de Vera (2010) modified this method by proposing an additive semiparametric model for clustered count data. The parametric part of the model takes care of the inherent clustering of observations due to demographic similarity or other spatial dependency mechanisms while the nonparametric part of the model makes the estimation of the effect of covariates flexible by using smoothing splines. This paper proposes a modification of de Vera's (2010) work by adding a term in the parametric part of the model which is supposed to handle the zero-inflation of the outcome variable. The predictive performance of the proposed method is compared to the ordinary ZIP to determine the advantages and disadvantages of the postulated model.

Section 2 gives a brief discussion on modeling count data together with the issue of zero-inflation. Section 3 introduces the postulated model and the proposed methods in estimating it. Results of the simulation study are presented in Section 4 while conclusions are stated in Section 5.

2. Related Literature

Based on the discussion of Mouatassim and Ezzahid (2012), the statistical technique assumes that the dependent variable follows a Poisson distribution

and the logarithm of its mean is modelled by a linear combination of covariates. Suppose Y_i is the response variable which has a Poisson distribution with parameter λ_i defined by as a function of the covariate $\underline{x}$. Its probability distribution is given by $P(Y_i = y_i) = \frac{e^{-\lambda_i}(\lambda_i)^{y_i}}{y_i!}$ whose the conditional mean is specified as $\lambda_i = E(Y_i \mid \underline{x}_i) = \exp(\underline{x}' \underline{\beta})$ where $\underline{x}_i' = (x_{i1}, x_{i2}, ..., x_p)'$ contains the explanatory variables and $\underline{\beta}' = (\beta_1, \beta_2, ..., \beta_p)'$ is the vector parameters. To estimate the parameters of the Poisson regression model, the maximum likelihood techniques may be used. This technique, however, assumes that the mean and the variance of the response variable are the same. Unfortunately, this assumption is usually violated in practice resulting in overdispersion of the data.

One of the sources of overdispersion is the excess of zeros in the values of the response variable. In this case, the Zero-Inflated Poisson (ZIP) regression model yields a better fit in the count data (Mouatassim and Ezzahid, 2012). Lambert (1992, as cited in Monod, 2011) introduced the model which started a class of regression model for excess zeros. Lee et al. (2011) described this model as a mixture of a discrete point mass and a Poisson distribution. The discrete point pass is the logistic portion of the model which models the probability of a count being zero. On the other hand, the Poisson portion is actually a zero-truncated Poisson distribution for nonzero values. Therefore, the ZIP model assumes that the outcomes come from two process. One handles the zero inflation by including a proportion $1- p$ of excess zero and a proportion of $p \exp(-\lambda_i)$ of zeros coming from the Poisson distribution while the second process models the positive counts using zero-truncated model. The ZIP model is given by

$$P(Y_i = y_i \mid \underline{x}_i, \underline{z}_i) = \begin{cases} \theta_i(\underline{z}_i) + (1-\theta_i(\underline{z}_i))Pois(\lambda_i, 0 \mid \underline{x}_i) & if\ y_i = 0 \\ (1-\theta_i(\underline{z}_i))Pois(\lambda_i, y_i \mid \underline{x}_i) & if\ y_i > 0 \end{cases}$$

where $\underline{z}_i$ is the vector containing the explanatory variable for the probability θ_i, Pois $(\lambda_i, 0 \mid \underline{x}_i) = \exp(-\lambda_i)$ and $Pois(\lambda_i, y_i \mid \underline{x}_i) = \frac{e^{-\lambda_i}(\lambda_i)^{y_i}}{y_i!}$. According to Lambert (1992, as cited in Mouatassim and Ezzahid, 2012), $\theta_i(\underline{z}_i)$ is modeled using Logit model given by:

$$\theta_i(\underline{z}_i) = \frac{\exp(\underline{z}_i' \underline{\gamma})}{1 + \exp(\underline{z}_i' \underline{\gamma})}$$

where $\underline{z}_i$ is the vector containing the covariates defining the probability θ_i and $\underline{\gamma}$ is the vector of parameters. Mouatassim and Ezzahid (2012) mentioned that the vector $\underline{z}_i$ can contain elements of $\underline{x}_i$ and Probit model can be used instead of the

logit model in modeling $\theta_i(\underline{z}_i)$. Furthermore, they noted that the parameter θ_i can be related to $\underline{\lambda}_i$ or completely independent with each other.

Hastie and Tibshirani (1990) introduced generalized additive model (GAM) which extends the concept of generalized linear model (GLM).These models are nonparametric generalization of the linear model; thus, nothing is imposed on the form of the dependency of the response variable Y on the predictors X_1, X_2, ..., X_p. Moreover, because they are additive in the predictor effects, they preserve the important feature of a model which is interpretability. These models can also guide the analyst in figuring out the appropriate shape of each of the predictor effect. Hastie and Tibshirani (1990) defined an additive model as follows:

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

where the error terms ε are independent of the X_j's, $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. The f_j's are arbitrary univariate functions, one for each explanatory variable. The f_j's can be viewed as smooth functions which are somehow can be individually estimated by a scatterplot smoother in an iterative manner.

Barry and Welsh (2002) used GAM in modeling zero-inflated count data with primary interest on the association between the presence and absence of a species and the available explanatory variables, and, conditional on the organism being present, on the correlation of abundance and the covariates. They derived the link and variance functions necessary in using GAM with zero-inflated data.

However, Barry and Welsh (2002) did not consider the clustering of data which exhibit spatial dependencies within a cluster and possibly independent across different clusters. de Vera (2010) mentioned that clustering of data may cause values of intercepts and coefficients of the predictors of the outcome variable vary. Additionally, Zyzanski et al. (2004, as cited in de Vera, 2010) remarked that when the intracluster correlation is ignored in the analysis, it could lead to invalid test results, and biased estimates and effect sizes. To analyze clustered data, Hedeker et al. (1994, as cited in de Vera, 2010) used a random effects regression model to assess the degree of spatial dependency of individuals within communities. Hall (2000, as cited in Lee et al., 2011) suggested a zero-inflated Poisson model with random effects for within-subject dependence but this was not used in the zero inflation part of the model.

This paper explores the ability of the proposed semiparametric (combination of parametric and nonparametric elements) regression model for clustered count data in terms of its predictive power.

3. Methodology

This section discusses the postulated model in handling zero-inflated response variable. It is followed by the estimation procedures and simulation scenarios conducted in this study.

3.1 Postulated model

This paper modified the model proposed by de Vera (2010) by adding the term λ_i in the model which handle the inflation of zeros in the response variable Y. The postulated model is given by

$$\log[E(Y_i^k \mid \mu_k)] = \mu_k + \lambda_i + \sum_{i=1}^{p} f(X_{ij}^k) \tag{1}$$

where

i = 1, 2, ..., n_k $\quad$ $j = 1, 2, ..., p$ $\quad$ $k = 1, 2, ..., n$

n_k is the cluster size

n is the total number of clusters

Y_i^k is the i^{th} value of the response variable within the k^{th} cluster

μ_k is a random cluster-specific intercept

$\lambda_i = \begin{cases} \lambda & \text{if } Y_i^k = 0 \\ 0 & otherwise \end{cases}$ is the term that will adjust the value of the response variable for excess zeros.

X_{ij}^k is the value of the j^{th} explanatory variable of the i^{th} observation within the k^{th} cluster

$f(X_{ij}^k)$ is the smoothed function of X_{ij}^k (nonparametric)

It is assumed that Y_i^k is a zero-inflated count response variable. The predictors X_{ij}^k can either be quantitative or qualitative. The random cluster-specific intercept μ_k will be used to take into account the clustering of the observations. Moreover, the clusters are assumed to be spatially independent of one another.

The classical backfitting algorithm is invoked in the estimation of the parametric and nonparametric portions of the postulated model in (1). Hastie and Tibshirani (1990) outlined the algorithm as follows:

(1) Initialize: $\alpha = ave(Y_i)$, $f_i = f_j^0$, $j = 1,2,...,p$.

(2) Cycle: $j = 1, 2, \ldots, p$

$$f_j = S_j[(Y - \alpha - \sum_{k \neq j} f_k \mid X_j],$$

where S_j $(Y \mid X_j)$ denotes a smooth of the response variable Y against the predictor X_j, and produces a function.

(3) Continue (2) until the individual functions do not change much.

The two methods proposed by de Vera (2010) are used in this paper. In Method 1 (Semiparametric estimation), the parametric and nonparametric parts are simultaneously estimated in a semiparametric sense. GAM fits the parametric linear model to account for the cluster effect μ_k, taking advantage of the inherent homogeneity within clusters, and zero-inflation term λ_i, to address the excess zeros in the response variable while a smoothing spline is used to estimate the nonparametric function $f(X_{ij}^k)$ inducing flexibility in the function by minimizing the penalized residual sum of squares given by

$$S(f) = \sum_{i=1}^{n} [Y_i - f(X_i)]^2 + \int_a^b \eta [f'(X)]^2 dx \qquad (2)$$

where $\eta > 0$ is the smoothing parameter. η manages the balance between smoothness as reflected in the second term of (2) and goodness of fit measured by the first term in (2).

Meanwhile, in Method 2 (Backfitting of Semiparametric Model), the nonparametric portion is estimated first, and then the parametric part is estimated from the residuals. As illustrated by de Vera (2010), $f(X_{i1}^k)$ is estimated nonparametrically using smoothing splines then the partial residuals $e_i^{(1)} = Y - \exp\{f(\hat{X}_{i1}^k)\}$. The partial residual $e_i^{(1)}$ contains information in Y that cannot be explained by $f(X_{i1}^k)$. Afterwards, $f(X_{i2}^k)$ is estimated this time with the partial residual $e_i^{(1)}$ as the new response variable. Then, the second partial residual $e_i^{(2)}$ is computed by $e_i^{(2)} = \hat{e}_i^{(1)} - \exp\{f(X_{i2}^k)\}$. The process continues until $f(X_{ip}^k)$ is estimated and the p^{th} partial residuals $e_i^{(p)}$ is computed. These partial residuals contain information about the zero-inflation of the response variables λ_i and cluster-specific random effect μ_k, that will be estimated parametrically. The predicted value of the response variable Y_i^k is computed by obtaining the sum of the estimates of the parameters and the nonparametric function. De Vera (2010) mentioned that Method 2 is more advantageous than Method 1 because it is more computationally simpler because each component is estimated one at a time. She also noted that thin plane smoothing splines should be used once there are two or more covariates involved whose convergence rate decreases as the number of covariates increases.

3.2. Simulation design

A simulated study is conducted in order to determine the performance of the proposed methodology. The distribution of the cluster-specific random effect, μ_k,

is varied by assuming a Normal and a Poisson distribution whose means differ across clusters. Moreover, to investigate the behaviour of the proposed method with respect to cluster sizes (small, medium, large) and number of clusters, these are also varied. It is interesting to find out the effect of the cluster size and number of clusters to the predictive accuracy of the proposed method. Cases which involve equal and unequal cluster sizes are also considered in the simulation procedure to know if the behaviour of the proposed methods differ.

Distribution of the explanatory variables is also assumed to be Normal or Uniform. There are two predictors used in the simulations. The generation of a zero-inflated Poisson response variable is based on SAS codes used by Erdman et al. (2008) with modifications so that observations within each cluster share a common characteristic. The process is replicated 100 times.

Table 3.2.1. Simulation Scenarios

1. distribution of the cluster specific effect (μ_k)	(a) $\mu_k \sim N(0.15k, 2)$ (b) $\mu_k \sim Po(0.15k)$
2. number of clusters (n)	Small – 3 clusters Medium – 5 clusters Large – 10 clusters
3. cluster size (n_k)	(a) Equal cluster sizes Small – 25 Medium – 50 Large – 100 (b) Unequal cluster sizes 3 clusters – 25, 50, 100 5 clusters – 25, 50, 75, 100, 125 10 clusters – 10, 25, 25, 50, 50, 75, 75, 100, 100, 120
4. distribution of X_{ij}^k	(a) $X_{ij}^k \sim N(0,1)$ (b) $X_{ij}^k \sim U(0,1)$
5. functional form of $f(X_{ij}^k)$	$f(X_{ij}^k) = \beta X_{ij}^k$
6. value of β in $f(X_{ij}^k)$	(a) $b_1 = 0.2$, $b_2 = 0.2$ (b) $b_1 = 0.4$, $b_2 = 0.4$

These simulations are conducted to compare among the two semiparametric methods and the ordinary ZIP model. To measure the predictive power of the models being compared, the Root Mean Square Error (RMSE) is used as given in the formula below. The Mean Absolute Percentage Error (MAPE) is not utilized because of the zero values of the response variable.

$$RMSE = \sqrt{\frac{\sum_k \sum_i (Y_i^k - \hat{Y}_i^k)^2}{\sum_k n_k}}$$

4. Results and Discussion

The predictive accuracy of the proposed semiparametric Poisson regression model for zero-inflated clustered data is the primary interest of this study using Methods 1 and 2. This will be compared to the ordinary ZIP model.

Table 4.1 presents the average root mean square error (RMSE) assuming equal cluster sizes for the different scenarios in the simulation study. As the number of clusters increases, the average RMSE's of all the methods also increase. A large difference can be observed between 5 and 10 clusters. For all methods, the average RMSE's of 10 clusters are roughly twice that of 5 clusters. This means that, for zero-inflated clustered data, methods perform better for small number of clusters. Predictive performance of the methods deteriorates as the number of cluster increases.

Table 4.1. Average RMSE for Equal Cluster Sizes

		Semiparametric Method (Method 1)	Backfitting Method (Method 2)	Ordinary ZIP (Method 3)
No. of Clusters	3	25.33	33.72	39.38
	5	38.13	47.73	54.66
	10	83.63	101.05	105.50
Cluster Sizes	25	34.43	50.96	62.50
	50	49.07	60.80	65.59
	100	63.58	70.74	71.45
Distribution of Predictors X_1 and X_2	N(0,1)	54.60	64.42	67.70
	U(0,1)	43.46	57.25	65.33
Coefficients	b_1= 0.2, b_2= 0.2	46.75	57.27	59.05
	b_1= 0.4, b_2= 0.2	51.31	64.40	73.97
Distribution of Cluster Specific Intercept μ_k	N(0.15k,2)	84.23	105.62	116.38
	Po(0.15k)	13.83	16.05	16.65

It is remarkable that Method 1 wherein the parametric and nonparametric parts of the proposed model are simultaneously estimated, obtained the lowest average RMSE among the three methods. This differs from the results of de Vera (2010) where Method 2 yields a lower MAPE than Method 1. Among the three methods being compared, Method 3 or the ordinary ZIP regression model obtained the highest RMSE, or, in other words, has the poorest predictive performance.

When cluster sizes are varied, the average RMSE of Methods 2 and 3 are comparable except for the case where $n_k = 25$ (small cluster size). Overall, Method 1 outperforms the other two methods by having the lowest average RMSEs. It can be observed that as the cluster sizes become larger, the average RMSEs also blow up. One possible reason is that as the cluster size increases, the observations become more homogeneous due to zero inflation. Hence, having very few non-zero observations will yield a tremendous escalation of the RMSE.

The distributions of covariates X_1 and X_2 are varied to determine its effects on predictive accuracy of the methods. As illustrated in Table 4.1, the average RMSEs of assuming N(0,1) are a little larger than those of U(0,1). This is, perhaps, due to larger variability of N(0,1) compared to U(0,1). This would mean that the more variable the predictors are, the predictive performance would be poorer. Among the three methods being assessed, Method 1 still has the lowest average RMSE.

To evaluate how the relative importance of covariates affect the predictive power of the postulated model, their coefficients are varied. In the first scenario, the coefficients are the same, meaning they are equally important, while, in the second, the first covariate has more influence than the second. In Table 4.1, the average RMSE of the first scenario is lower than that of the second. It is possible that because the first covariate has a larger coefficient, it adds to the differences on the values of the zero-inflated response variable resulting to a more difficult to predict outcome variable. Moreover, Method 1 performs best followed by Methods 2 and 3, respectively.

There are two distributional assumptions of μ_k considered in this simulation study – Normal and Poisson distributions. Noticeably, the average RMSEs when normal distribution is assumed are much larger than those of Poisson distribution. First, the variance of the normal distribution is fixed at 2 while the variance of the Poisson distribution is also varied. Based on the specifications, the variance of the Poisson distribution is smaller than that of the Normal; hence, producing very small RMSEs. Furthermore, Methods 2 and 3 are comparable in both scenarios. Method 1 still obtained the lowest RMSE, thus, highest predictive accuracy.

Meanwhile, Table 4.2 shows the average RMSEs assuming unequal cluster sizes as specified in Table 3.2.1. As the number of clusters increases, the average RMSE also increases for all methods. The average RMSE for 5 cluster case increases also by 50% when the number of clusters become 10. Thus, the predictive accuracy of the methods for zero-inflated clustered data declines when there are many clusters being considered. In addition, Methods 2 and 3 are comparable. Method 1 has the lowest RMSE for the three simulation scenarios.

Table 4.2. Average RMSE for Unequal Cluster Sizes

		Semiparametric Method (Method 1)	Backfitting Method (Method 2)	Ordinary ZIP (Method 3)
No. of Clusters	3	27.56	31.33	35.27
	5	45.22	48.39	49.38
	10	97.84	103.88	104.78
Distribution of Predictors X_1 and X_2	N(0,1)	62.13	66.26	66.06
	U(0,1)	51.62	56.14	60.22
Coefficients	$b_1 = 0.2$, $b_2 = 0.2$	55.05	58.91	61.79
	$b_1 = 0.4$, $b_2 = 0.2$	58.70	63.49	64.50
Distribution of Cluster Specific Intercept μ_k	N(0.15k,2)	93.24	100.36	104.07
	Po(0.15k)	20.51	22.04	22.21

When the distributional assumption of the covariates is considered, N(0,1) yields larger RMSE compared to U(0,1) primarily because of higher degree of heterogeneity. Moreover, for the N(0,1) case, Methods 2 and 3 are comparable, this is not the case for U(0,1). Method 1 performs best regardless of the distributional assumption of the covariates.

As the relative importance of the covariates is varied, it can be observed that the average RMSEs of the two scenarios are close to one another which means that the three methods are fairly robust to the degree of importance of the covariates when cluster sizes are unequal. Method 1 yield the best predictive accuracy among the three methods in this case.

As shown in Table 4.2, RMSE are larger when cluster-specific random intercept μ_k is normal compared to Poisson. This is due to larger variability induced when normal assumption is imposed on μ_k. The average RMSEs of Methods 2 and 3 are comparable while that of Method 1 bested the other two.

5. Conclusions

The postulated additive semiparametric Poisson regression model for zero-inflated clustered data in (1) given n clusters with n_k observations is given by $\log[E(Y_i^k \mid \mu_k)] = \mu_k + \lambda_i + \sum_{i=1}^{p} f(X_{ij}^k)$ where μ_k is the cluster-specific random intercept, and λ_i accounts for zero-inflation of the outcome variable. To estimate the postulated model, two methods based on de Vera (2010) are considered. The first estimates both the parametric and nonparametric parts of the model simultaneously, while the second method utilizes the backfitting algorithm taking advantage of the additive form of the model. In the simulation study, the first method outperforms the second and the ordinary ZIP regression model when it comes to predictive performance whether the cluster sizes are equal or not. In fact, the average RMSE of the ordinary ZIP model is at most 81% and 27% higher than that of the Method 1 for the balanced and unbalanced cases, respectively, i.e. equal and unequal cluster sizes.

Furthermore, as the number of clusters increases and as the cluster sizes becomes larger, the predictive ability of the three methods deteriorates tremendously. Hence, Method 1 performs best for small number of clusters with few numbers of observations.

REFERENCES

BAGHBAN, A., POURHOSEINGHOLI, A., ZAYERI, F. ASHTARI, S. and ZALI, M., 2013, Zero inflated statistical count models for analyzing the costs imposed by GERD and dyspepsia, *Arab Journal of Gastroenterology* 14:165–168.

BARRY, S. and WELSH, A., 2002, Generalized additive modelling and zero inflated count data, *Ecological Modelling* 157:179-188.

DE VERA, E., 2010, Semiparametric Poisson Regression Model for Clustered Data, Unpublished Thesis, School of Statistics, University of the Philippines.

DEMIDENKO, E., 2007, Poisson regression for clustered data, *International Statistical Review* 75:96-113.

ERDMAN, D., JACKSON, L., and SINKO, A., 2008, Zero-inflated poisson and zero-inflated negative binomial models using the COUNTREG procedure (Paper 322-2008), SAS Institute Inc., Cary, NC. Available at http://www2.sas.com/proceedings/forum2008/322-2008.pdf.

HASTIE, T. and TIBSHIRANI, R., 1990, *Generalized Additive Models*, London: Chapman and Hall.

JANG, H., LEE, S., and KIM, S., 2010, Bayesian analysis for zero-inflated regression models with the power prior: Applications to road safety countermeasures, *Accident Analysis and Prevention* 42: 540–547.

LEE, K., JOO, Y., SONG, J. and HARPER, D., 2011, Analysis of zero-inflated clustered count data: A marginalized model approach, *Computational Statistics and Data Analysis* 55:824-837.

LEE, Y, LEE, J., SONG, Y. and LEE, S., 2007, An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST, *Scientometrics* 70 (1):27-39.

MONOD, A., 2011, Generalized estimating equations for zero-inflated spatial count data, *Procedia Environmetal Sciences* 7:281-286.

MOUATASSUM, Y. and EZZAHID, E., 2012, Poisson regression and Zero-inflated Poisson regression: Application to private health insurance data, *European Actuarial Journal* 2:187-204.

NAYA, H., URIOSTE, J., CHANG, Y., MOTTA, M., KREMER, R., and GIANOLA, D., 2008, A comparison between Poisson and zero-inflated Poisson regression models with an application to number of black spots in Corriedale sheep, *Genet. Sel. Evol* 40:379-394.

ACKNOWLEDGMENT

The author would like to thank Dr. Erniel Barrios, Dr. Joseph Ryan Lansangan, and Ms. Iris Ivy Gauran for their insights and comments leading to the improvement of this paper.