Online Supplementary Material for

"Review and Recommendations for Zero-inflated Count Regression Modeling of Dental Caries Indices in Epidemiological Studies"

by J.S. Preisser, J. W. Stamm, D. L. Long and M. E. Kincade

Introduction

The purpose of this supplementary material is to provide examples and further discussion to support one of the main results of the article by Preisser, Stamm, Long and Kincade [2012]. The result is the claim that use of the at-risk latent class incidence rate ratio, or IRR-latent, as an estimate of the IRR for caries severity in the overall population, or IRR-severity, will *usually* result in bias towards the null hypothesis of no risk factor effect in the sense that the IRR-latent will be closer to 1.0 than IRR-severity. The result of the IRR-latent being "biased towards the null" almost always corresponds to γ_1 and β_1 in equation 1 of Preisser *et al.* [2012] (or equation 3 below) having opposite signs, which indicates that the effects of a risk factor on the probability of having an excess zero and on the mean in the at-risk latent class, respectively, will have consistent trends. The scenario of consistent trends is where a covariate decreases (increases) the probability of an excess zero and increases (decreases) the at-risk class mean. When trends are consistent, IRR-latent will *almost always* be biased towards the null. Using real-life data, we find that consistent trends, given by "opposite signs" of the regression coefficients for a covariate, occur much more frequently than "same signs". Data from the Dunedin Multidisciplinary Health and Development Study previously analyzed by Lewsey and Thomson [2004] are used for illustration.

Methods

Lewsey & Thomson [2004] performed a series of cross-sectional analyses to investigate risk factors for dental caries. Results of their fitted ZINB models that include covariates in both model parts are shown in Table 1. For convenience, we define the following notation for covariates: x_{i1} = SESLOW, x_{i2} = SESMED, x_{i3} = FEXP $< 50\%$, and x_{i4} = FEMALE. Note that SESLOW (low SES) and SESMED (medium SES) are two dummy variables for socioeconomic status with "high SES" being the reference category. Note that x_{i1} and x_{i2}

Table 1: The parameter estimates and 95% confidence intervals for ZINB models in equations (1) and (2) from Table 1 of Lewsey & Thomson [2004]

		Logit for excess zeros		Negative Binomial
		estimate (95% CI)		estimate (95% CI)
Age 5 years (dmfs)				
Intercept	γ_0	-0.762 (-1.587, 0.063)	β_0	1.061 (0.675, 1.447)
SESLOW	γ_1	-0.508 (-1.394, 0.378)	β_1	0.861 (0.447, 1.275)
SESMED	γ_2	-0.320 (-1.061, 0.421)	β_2	0.484 (0.117, 0.851)
FEXP $<$ 50%	γ_3	0.024 (-0.490, 0.538)	β_3	-0.003 (-0.234, 0.228)
FEMALE	γ_4	-0.013 (-0.530, 0.504)	β_4	-0.085 (-0.318, 0.145)
Age 18 years (DMFS)				
Intercept	γ_0	-1.288 (-1.925, -0.651)	β_0	1.929 (1.743, 2.115)
SESLOW	γ_1	-1.304 (-2.284, -0.324)	β_1	0.226 (0.008, 0.444)
SESMED	γ_2	-1.014 (-1.659, -0.369)	β_2	0.144 (-0.040, 0.328)
FEXP $<$ 50%	γ_3	-0.245 (-0.835, 0.345)	β_3	0.089 (-0.033, 0.211)
FEMALE	γ_4	0.384 (-0.196, 0.964)	β_4	-0.085 (-0.318, 0.148)
Age 26 years (DMFS)				
Intercept	γ_0	-3.153 (-5.201, -1.105)	β_0	2.268 (2.080, 2.456)
SESLOW	γ_1	-0.983 (-3.343, 1.377)	β_1	0.397 (0.178, 0.618)
SESMED	γ_2	-0.532 (-2.110, 1.046)	β_2	0.205 (0.019, 0.391)
FEXP $<$ 50%	γ_3	-1.070 (-2.934, 0.794)	β_3	0.132 (0.005, 0.259)
FEMALE	γ_4	0.809 (-0.839, 2.457)	β_4	0.021 (-0.106, 0.148)

Reproduced with permission from *Community Dentistry & Oral Epidemiology* (pending).

together determine socioeconomic status with $x_{i1} = x_{i2} = 0$ for high SES. Dichotomous variables were also created for a person's residential fluoride exposure less than half their life (FEXP $<$ 50%, coded as 1) and FEMALE (yes, coded as 1). Thus, letting $\mathbf{x_i}$ denote the full set of covariates for the i-th person, the probability of an excess zero in the ZINB model is:

$$\psi_i(\mathbf{x_i}) = \frac{\exp(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4})}{1 + \exp(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4})} \tag{1}$$

and the mean dmfs for at-risk children is

$$\mu_i(\mathbf{x_i}) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}). \tag{2}$$

Thus, the overall mean dmfs for child i, or severity given by $\mu_i(\mathbf{x_i})(1 - \psi_i(\mathbf{x_i}))$, is:

$$\mathrm{E}(Y_i|\mathbf{x_i}) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4})}{[1 + \exp(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4})]}.$$

IRR (severity) is determined by taking the ratios of overall means, say $\mathrm{E}(Y_i|\mathbf{x_i} = \mathbf{A})$, for subjects in group A given by one set of covariates, to $\mathrm{E}(Y_i|\mathbf{x_i} = \mathbf{B})$, for subjects in group B given by another set. For example, the IRR (severity) for low SES relative to high SES, holding values of x_{i3} and x_{i4} fixed, is:

$$\frac{E(Y_i|x_{i1} = 1, x_{i2} = 0, x_{i3}, x_{i4})}{E(Y_i|x_{i1} = 0, x_{i2} = 0, x_{i3}, x_{i4})} = \exp(\beta_1)\frac{[1 + \exp(\gamma_0 + \gamma_3 x_{i3} + \gamma_4 x_{i4})]}{[1 + \exp(\gamma_0 + \gamma_1 + \gamma_3 x_{i3} + \gamma_4 x_{i4})]} \tag{3}$$

The following observations can be made with respect to this formula:

- The IRR-severity for a covariate, in this case SESLOW, depends on the values of the other covariates. For example, if $x_{i3} = 0$ and $x_{i4} = 0$,

  $$\frac{E(Y_i|x_{i1} = 1, x_{i2} = 0, x_{i3} = 0, x_{i4} = 0)}{E(Y_i|x_{i1} = 0, x_{i2} = 0, x_{i3} = 0, x_{i4} = 0)} = \exp(\beta_1)\frac{[1 + \exp(\gamma_0)]}{[1 + \exp(\gamma_0 + \gamma_1)]}$$

  which, for Age 5 years (Table 2), upon plugging in estimates is 2.709. Specifically, this is the IRR-severity comparing a male child of low socioeconomic status to a male child of high socioeconomic status, when both children have residual flouride exposure with half or more of their life. Now, if for both children we change gender to female (i.e., $x_{i4} = 1$) or flouride exposure to less than half of their life ($x_{i3} = 1$), we will get different estimates. But these estimates given by the different combinations of x_{i3} and x_{i4} vary only slightly for this example as shown in Table 2.

- A covariate-adjusted IRR-severity may be obtained by inserting mean values of covariates (in this case, proportions) for x_{i3} and x_{i4} into equation (3), obtaining

  $$\frac{\hat{E}(Y_i|x_{i1} = 1, x_{i2} = 0, x_{i3} = \bar{x}_3, x_{i4} = \bar{x}_4)}{\hat{E}(Y_i|x_{i1} = 0, x_{i2} = 0, x_{i3} = \bar{x}_3, x_{i4} = \bar{x}_4)} = \exp(\hat{\beta}_1)\frac{[1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_3\bar{x}_3 + \hat{\gamma}_4\bar{x}_4)]}{[1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 + \hat{\gamma}_3\bar{x}_3 + \hat{\gamma}_4\bar{x}_4)]} \tag{4}$$

- If only an intercept term is included in the excess zeros part of the model (e.g., $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0$ in equation (1)), then the IRR-latent and IRR-severity are the same. Thus, the IRR-severity for SESLOW versus high SES (the reference category) is

$\exp(\beta_1)$ for any covariate values x_{i3} and x_{i4} that are held in common for both children. Similarly, the IRR-severity for SESMED versus High SES is $\exp(\beta_2)$, and for SESLOW vs SESMED it is $\exp(\beta_1 - \beta_2)$. This equivalency of IRR-latent and IRR-severity also occurs for socioeconomic status when SES is not included in the excess zeros part of the model, and other covariates (x_{i3} or x_{i4}) are included in both model parts. Equivalency results because the ratio factor on the right-hand-side of equation (3) simplifies to 1.0.

For completeness, we define IRR-severity for medium relative to high SES:

$$\frac{E(Y_i|x_{i1}=0, x_{i2}=1, x_{i3}, x_{i4})}{E(Y_i|x_{i1}=0, x_{i2}=0, x_{i3}, x_{i4})} = \exp(\beta_2)\frac{[1+\exp(\gamma_0+\gamma_3 x_{i3}+\gamma_4 x_{i4})]}{[1+\exp(\gamma_0+\gamma_2+\gamma_3 x_{i3}+\gamma_4 x_{i4})]} \tag{5}$$

and low relative to medium SES (the remaining contrast for SES)

$$\frac{E(Y_i|x_{i1}=1, x_{i2}=0, x_{i3}, x_{i4})}{E(Y_i|x_{i1}=0, x_{i2}=1, x_{i3}, x_{i4})} = \exp(\beta_1-\beta_2)\frac{[1+\exp(\gamma_0+\gamma_2+\gamma_3 x_{i3}+\gamma_4 x_{i4})]}{[1+\exp(\gamma_0+\gamma_1+\gamma_3 x_{i3}+\gamma_4 x_{i4})]}. \tag{6}$$

Next, the IRR-severity for FEXP $<$ 50% is:

$$\frac{E(Y_i|x_{i1}, x_{i2}, x_{i3}=1, x_{i4})}{E(Y_i|x_{i1}, x_{i2}, x_{i3}=0, x_{i4})} = \exp(\beta_4)\frac{[1+\exp(\gamma_0+\gamma_1 x_{i1}+\gamma_2 x_{i2}+\gamma_4 x_{i4})]}{[1+\exp(\gamma_0+\gamma_1 x_{i1}+\gamma_2 x_{i2}+\gamma_3+\gamma_4 x_{i4})]} \tag{7}$$

and the IRR-severity for FEMALE is:

$$\frac{E(Y_i|x_{i1}, x_{i2}, x_{i3}, x_{i4}=1)}{E(Y_i|x_{i1}, x_{i2}, x_{i3}, x_{i4}=0)} = \exp(\beta_4)\frac{[1+\exp(\gamma_0+\gamma_1 x_{i1}+\gamma_2 x_{i2}+\gamma_3 x_{i3})]}{[1+\exp(\gamma_0+\gamma_1 x_{i1}+\gamma_2 x_{i2}+\gamma_3 x_{i3}+\gamma_4)]}. \tag{8}$$

The next section compares equations (3), (5), (6), (7) and (8) to their at-risk latent class IRR counterparts, identifying instances when covariates have consistent or inconsistent trends in the two ZI model parts. For example, we examine IRR-severity for FEMALE in order to illustrate the impact of the relatively uncommon case where the covariates have inconsistent trends (i.e. coefficient pairs with same signs in Table 1).

Table 2: Comparison of IRR estimates for the effect of socio-economic status (SES) in the at-risk latent class versus the IRR estimates for the effect of SES in the overall population based on ZINB model parameter estimates reported in Table 1

		Low vs High		Med vs High		Low vs Med	
x_{i3}	x_{i4}	Latent	overall	Latent	overall	Latent	overall
		Age 5 years (dmfs)					
0	0	2.366	2.709	1.623	1.777	1.458	1.524
0	1	2.366	2.705	1.623	1.776	1.458	1.523
1	0	2.366	2.715	1.623	1.780	1.458	1.525
1	1	2.366	2.712	1.623	1.779	1.458	1.525
		Age 18 years (DMFS)					
0	0	1.254	1.488	1.155	1.339	1.086	1.111
0	1	1.254	1.587	1.155	1.415	1.086	1.122
1	0	1.254	1.440	1.155	1.302	1.086	1.106
1	1	1.254	1.520	1.155	1.364	1.086	1.114
		Age 26 years (DMFS)					
0	0	1.487	1.527	1.228	1.249	1.212	1.223
0	1	1.487	1.574	1.228	1.274	1.212	1.236
1	0	1.487	1.501	1.228	1.235	1.212	1.215
1	1	1.487	1.518	1.228	1.244	1.212	1.212

Results

The results in Tables 2, 3 and 4 provide estimates of IRR-severity for the overall population that are conditional on the values of the other covariates. It is worth noting that with descriptive data on overall means (in this case, proportions of SESLOW, SESMED, FEXP $< 50\%$ and FEMALE, not provided by Lewsey and Thompson), a marginal covariate-adjusted IRR-severity could, in principle, be computed as in equation (4).

In 10 out of 12 cases in Table 1 (intercept terms excluded), $\hat{\gamma}_1$ and $\hat{\beta}_1$ have opposite signs revealing consistent trends for the risk factor on the probability of an excess zero and the at-risk latent class mean. Through this example, we seek to confirm the expectation that the use of the at-risk latent class IRR for dental caries as an estimate of IRR-severity in the overall population will *usually* result in bias towards the null hypothesis of no risk factor effect in the sense that the IRR-latent will be closer to 1.0 than IRR-severity. Recall that, because SESLOW and SESMED are included in the excess zeros part of the model, the values of IRR-severity for SES will vary by the other covariates in the model, as shown in equation (3). We make the following observations about socioeconomic status from the

Table 3: Comparison of IRR estimates for the effect of FEXP $< 50\%$ versus FEXP $\geq 50\%$ in the at-risk latent class versus the IRR estimates for the effect of gender in the overall population based on ZINB model parameter estimates reported in Table 1

		Age 5 years		Age 18 years		Age 26 years	
SES	gender	Latent	overall	Latent	overall	Latent	overall
LOW	male	0.997	0.992	1.093	1.110	1.141	1.153
	female	0.997	0.992	1.093	1.117	1.141	1.168
MED	male	0.997	0.991	1.093	1.115	1.141	1.160
	female	0.997	0.991	1.093	1.124	1.141	1.183
HIGH	male	0.997	0.989	1.093	1.147	1.141	1.173
	female	0.997	0.989	1.093	1.166	1.141	1.211

results in Table 2:

- Because $\hat{\gamma}_1 < 0$, $\hat{\gamma}_2 < 0$ and $\hat{\gamma}_1 < \hat{\gamma}_2$ for each age group (Table 1), the IRR-latent estimates $\exp(\hat{\beta}_1)$, $\exp(\hat{\beta}_2)$ and $\exp(\hat{\beta}_1 - \hat{\beta}_2)$ are smaller than (underestimate) the IRR-severity estimates given in equations (3), (5) and (6), respectively. Furthermore, since the signs of the $\gamma-$coefficients are opposite of the signs of their respective $\beta-$coefficients, the SES risk factors show consistent trends on the probability of an excess zero and on the at-risk latent class means. Thus, according to expectations, the IRR-latent estimates for SES have values closer to 1.0 than IRR-severity estimates, i.e., they are biased towards the null hypothesis of no SES effect.

- For age 5, the IRR-severity estimates are practically the same across the four groups given by fluoride exposure and gender. The similarity of estimates arises because the estimates for FEXP and FEMALE in the excess zero part of the model (i.e., $\hat{\gamma}_3$ and $\hat{\gamma}_4$ in Table (1)) are practically equal to zero. For ages 18 and 26, $\hat{\gamma}_3$ and $\hat{\gamma}_4$ are farther from zero, so there is more variation in the estimates of IRR-severity for SES among the rows of Table 2.

- The main finding is that there are many non-negligible differences between IRR-latent and IRR-severity for SES in Table 2.

For FEXP, IRR-latent estimates are always closer to 1.0 (i.e., no effect) than IRR-severity estimates (Table 3), although differences are small. This meets expectations because

Table 4: Comparison of IRR estimates for the effect of FEMALE versus MALE in the at-risk latent class versus the IRR estimates for the effect of gender in the overall population based on ZINB model parameter estimates reported in Table 1

		Age 5 years		Age 18 years		Age 26 years	
SES	FEXP	Latent	overall	Latent	overall	Latent	overall
LOW	$\geq 50\%$	0.919	0.921	0.919	0.890	1.021	1.002
	$< 50\%$	0.919	0.921	0.919	0.895	1.021	1.014
MED	$\geq 50\%$	0.919	0.922	0.919	0.881	1.021	0.991
	$< 50\%$	0.919	0.922	0.919	0.888	1.021	1.011
HIGH	$\geq 50\%$	0.919	0.922	0.919	0.834	1.021	0.972
	$< 50\%$	0.919	0.922	0.919	0.848	1.021	1.003

$\hat{\gamma}_3$ and $\hat{\beta}_3$ have opposite signs for each age (Table 1). For age 5 years, IRR-latent estimates are greater than IRR-severity estimates because $\hat{\gamma}_3 > 0$. Also, $\hat{\gamma}_3$ and $\hat{\beta}_3$ are close to zero, so the IRR-latent and IRR-severity estimates are close to 1.0. For ages 18 and 26, $\hat{\gamma}_3 < 0$ resulting in the IRR-latent being smaller than IRR-severity. There is mild variation among the IRR-severity estimates for FEXP across the rows of Table 3 for ages 18 and 26 years.

Inconsistent trends (same signs) for effects on the probability of an excess zero and the at-risk latent class mean were seen for two of three ages (5 and 26 years) for FEMALE. For age 5, $\hat{\gamma}_4 = -0.013$ has a negative sign (Table 1) so IRR-latent underestimates IRR-severity. Specifically, Table 4 shows that the at-risk latent class IRR for FEMALE is $\exp(-0.085) = 0.919$ which is smaller than the IRR-severity since $\hat{\gamma}_4 < 0$ (see equation (8)). At the same time, $\hat{\beta}_4 = -0.085$ and $\hat{\gamma}_4 = -0.013$ have inconsistent trends (same signs) so that IRR-latent for FEMALE (age 5) is biased away from the null. Conversely, for age 18, the bias of IRR-latent for FEMALE is towards the null, because, in this case, $\hat{\beta}_4$ and $\hat{\gamma}_4$ have opposite signs. Also, since $\hat{\gamma}_4 > 0$, the IRR-latent is greater than the IRR-severity.

Finally, for age 26, both $\hat{\beta}_4 = 0.809$ and $\hat{\gamma}_4 = 0.021$ have positive signs (Table 1). Table 4 shows that the at-risk latent class IRR for FEMALE is $\exp(0.021) = 1.021$ which is greater than the IRR-severity since $\hat{\gamma}_4 > 0$. Moreover, since trends are inconsistent, the bias of IRR-latent for FEMALE is away from the null, but only in five of six cases for age 26. For adults with high SES and FEXP $> 50\%$, the IRR-severity estimate of 0.972 in the fifth row of Table 4 is actually farther from 1.0 than the IRR-latent estimate of 1.021. This anomaly may happen when the estimates for IRR-severity (> 1) and IRR-latent (< 1) are in the opposite direction. Such an anomaly, defying the laws for the impact on bias of consistent and inconsistent trends discussed in Preisser *et al.* [2012], occurred only once out of 72 times (1.4%) across all results in the three tables.

In summary, in most cases, the latent class IRR estimates tended to underestimate the IRR-severity estimates in the sense of having values closer to 1.0. These differences were most pronounced for SES, where the effects were largest.

Conclusions

As anticipated, the at-risk latent class IRR estimates tended to underestimate the IRR effects in the overall population in the sense of having values closer to 1.0. The explanation for this direction of bias is the consistent trends for the effects of caries risk factors on the probability of an excess zero and on the at-risk latent class mean demonstrated in 10 of 12 coefficient pairs (Table 1). If we additionally consider ZINB model results reported in Table 2 of Lewsey and Thomson [2004], Table 4 of Hashim *et al.* [2006], and Table 1 of Solinas *et al.* [2009], then 74% (26/35) of coefficient pairs in ZINB models with covariates in both model parts had consistent trends. Moreover, in all nine instances of inconsistent trends the coefficient for the covariate in the excess zero part of the model was not statistically significant at the 0.05 level. This suggests true inconsistent trends may be rare. Thus, the at-risk latent class IRR estimates will almost always be biased towards the null hypothesis of no covariate effect in large samples. While apparent bias was often small in the Dunedin study data analyzed by Lewsey and Thomson [2004] and in this manuscript of supplementary material, differences of mild and moderate size between IRR-latent and IRR-severity were also observed. For example, for the SESLOW covariate from Table 1 and Table 2, bias was a large as 21% (Age 18 years (DMFS), row in Table 2 with $x_{i3} = 0$ and $x_{i4} = 1$ where bias $= (1.587 - 1.254)/1.587 \times 100\% = 21\%$). Therefore, it is suggested that the latent class estimates not be considered as substitutes or proxies for properly computed "overall" estimates of severity.

Whether the misleading/incorrect/imprecise interpretations in the articles reviewed by Preisser *et al.* [2012] led the authors of those papers to unsubstantiated conclusions is difficult to answer based upon the results reported in those published papers. Only four of the eight articles reviewed that included covariates in both model parts of ZIP or ZINB models reported the full set of estimates of model coefficients (Campus *et al.* [2009] report similar results as Solinas *et al.* 2009 and were not included in the tally reported in the previous paragraph). Even when IRR-severity estimates can be computed, the assessment of uncertainty, whether via reported standard errors, confidence intervals or results of hypothesis testing, is necessary to determine whether or not researchers' conclusions would have changed had they estimated IRR-severity. When covariates are included in both model parts, the covariances in addition to variances of the regression coefficient estimates are needed to determine the uncertainty associated with estimates of overall effects such as those in equations (3), (5), (6), (7) and (8). Then, large sample variances of IRR-severity can be determined, via application of the delta method. Variances are then estimated by plugging in the appropriate maximum likelihood estimates of regression coefficients. The ultimate question of whether the misleading/incorrect/imprecise interpretations led the authors of the papers reviewed by Preisser *et al.* [2012] to unsubstantiated conclusions must be left to future research.

References

1. Campus G, Solinas G, Strohmenger L, Cagetti MG, Senna A, Minelli L, Majori S, Montagna MT, Reali D, Castiglia P. National pathfinder survey on children's oral health in Italy: pattern and severity of caries disease in 4-year-olds. Caries Research 2009; 43: 155-62.

2. Hashim R, Thomson WM, Ayers KMS, Lewsey JD, Awad M. Dental caries experience and use of dental services among preschool children in Ajman, UAE. International Journal of Paediatric Dentistry 2006; **16**, 257262.

3. Lewsey J, Thomson W. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. Community Dentistry and Oral Epidemiology 2004; **32**, 18389.

4. Preisser JS, Stamm JW, Long DL, Kincade ME. Review and Recommendations for Zero-inflated Count Regression Modeling of Dental Caries Indices in Epidemiological Studies. In Revision, Caries Research 2012.

5. Solinas G, Campus G, Maida C, Sotgiu G, Cagetti MG, Lesaffre E, Castiglia P. What statistical method should be used to evaluate risk factors associated with dmfs index? Evidence from the National Pathfinder Survey of 4-year-old Italian children. Community Dent Oral Epidemiol 2009; **37**, 539-546.