Web-based Supplementary Materials for "Marginal Mean Models for Zero-Inflated Count Data" by D. Todem, W-W Hsu and KM Kim

Web Appendix A: Additional simulation studies

We conduct an additional simulation study to evaluate the finite sample performances of the derived and the direct estimates of covariates effects on the marginal mean when a second order Taylor expansion of $\log\{\pi_i\}$ is invoked. For this simulation, we continue to assume that the exposure of interest is binary v_{1i} and the confounder v_{2i} is continuous. The data are generated from a ZINB with dispersion parameter $\kappa = 0.5$, and the true models for μ_i and π_i given by $\log\{\mu_i\} = 1.5-0.5v_{1i}-0.1v_{2i}$ and $\text{logit}\{\pi_i\} = 1.5-0.5v_{1i}-0.2v_{2i}$. The working regression models for μ_i and π_i are $\log\{\mu_i\} = \alpha_0+\alpha_1 v_{1i}+\alpha_2 v_{2i}$ and $\text{logit}\{\pi_i\} = \gamma_0+\gamma_1 v_{1i}+\gamma_2 v_{2i}$. From these working models, the MR estimate for the derived approach was computed assuming a second order Taylor expansion of $\log\{\pi_i\}$. This expansion includes the terms v_{1i}, v_{2i}, v_{2i}^2, and $v_{1i}v_{2i}$. We also used these terms to fit the marginal log-logit model with mean $\log\{\text{E}(\text{Y}_i|\text{V}_i)\} = \beta_0 + \beta_1 v_{1i} + \beta_2 v_{2i} + \beta_3 v_{2i}^2 + \beta_4 v_{1i}v_{2i}$. Estimates of the true MR are $\exp\{\hat{\beta}_{1,der} + 0.5\hat{\beta}^2_{3,der}\}$ and $\exp\{\hat{\beta}_{1,dir} + 0.5\hat{\beta}^2_{3,dir}\}$ for the derived and the direct methods. Finally, all simulations were replicated 500 times and for sample sizes 100, 400 and 1000.

Table W.1: Simulation results for the mean ratio (exposed vs unexposed), using the derived estimations from a conditional (conventional) log-logit model, and the direct estimation from a marginal log-logit working model, data generated from ZINB models, 500 resamples

		Conditional log-logit working model				Marginal log-logit working model			
		Derived marginal estimation				Direct marginal estimation			
True MR	$n^{\mp}$	EMR	% bias	MSE	CP(%)	EMR	% bias	MSE	CP(%)
0.543	100	0.569	4.800	0.021	96.7	0.584	7.473	0.025	94.1
	400	0.546	0.541	0.004	>99.9	0.549	1.103	0.004	93.4
	1000	0.542	-0.154	0.002	>99.9	0.544	0.102	0.002	92.4

$^{\mp}n/2$ is sample size per group; Data generated from ZINB model with a conditionally specified mean.

Results in Table W.1 show that when higher order variables are added in the design matrix for the marginal mean, we continue to have satisfactory results in that the bias on the mean ratio between the exposed and the unexposed groups is very negligible for the derived and direct estimates of the true mean ratio. Moreover the average MSE also decreases with increasing sample sizes leading to a conjecture that the invoked estimates are consistent. Most importantly, the direct estimation approach based on $\widehat{\beta}_{dir}$ continue to have appropriate coverage probabilities at the nominal 95% level, despite the fact that additional variables are added in the model. However, the derived estimation approach continues to exhibit 95% coverage probabilities higher than the nominal level.

Web Appendix B: Additional simulation studies

Finally, we conduct a simulation study to evaluate the accuracy of the Taylor approximation when a first order Taylor expansion of $\log\{\pi_i\}$ is assumed. More specifically, we compare the derived marginal mean on the log scale $\widehat{\beta}'_{der}\mathrm{X}_i$ with $\mathrm{X}_i = \mathrm{V}_i = (1, v_{1i}, v_{2i})'$ to the log of the fitted mean $\pi_i(\widehat{\gamma})\mu_i(\widehat{\alpha})$ from the fitted latent class regression models for π_i and μ_i, using Pearson and the Spearman correlations. In this study, we assume the same setting as in the paper with the exposure of interest being binary v_{1i} and the confounder v_{2i} continuous. Results in Table W.2 show satisfactory results for the approximation with average Person and Spearman correlations between $\log\{\pi_i(\widehat{\gamma})\mu_i(\widehat{\alpha})\}$ and $\widehat{\beta}'_{der}\mathrm{X}_i$ being above 97%.

Table W.2: Simulation results comparing $\log\{\pi_i(\widehat{\gamma})\mu_i(\widehat{\alpha})\}$ and $\widehat{\beta}'_{der}\mathrm{X}_i$ with $\mathrm{X}_i = \mathrm{V}_i = (1, v_{1i}, v_{2i})'$ from working latent class regression models for π_i and μ_i, using Pearson and the Spearman correlations with data were generated from ZINB models with marginally and conditionally specified means

	ZINB model with a conditionally specified mean		ZINB model with a marginally specified mean	
$n^{\mp}$	Spearman	Pearson	Spearman	Pearson
100	0.974	0.983	0.979	0.993
200	0.993	0.996	0.995	0.999
400	>0.999	>0.999	0.997	>0.999
1000	>0.999	>0.999	>0.999	>0.999

$^{\mp}n/2$ is sample size per group

Web Appendix C: Additional analysis of the Detroit Caries indices in primary dentition

Table W.3: Parameter estimates, standard errors (SE) and p-values for ZINB model under the marginal log-logit models with higher order covariate terms, consistent with a 2nd order Taylor expansion

Effects	Estimate	SE	p-value
Mean			
Intercept	1.4862	0.1432	$<.0001$
Age	0.8760	0.2035	$<.0001$
Unemployed	0.1989	0.1558	0.2020
PHP	0.006071	0.1020	0.9525
SI	0.3052	0.1635	0.0623
Age*SI	-0.08310	0.2441	0.7336
Unemployed*SI	-0.1546	0.1636	0.3451
PHP*SI	-0.02824	0.1046	0.7873
SI*SI	-0.07807	0.05136	0.1289
Age*SI*SI	0.05988	0.1240	0.6292
Age*Age	-0.3505	0.1136	0.0021
Unemployed*Age	0.2170	0.1524	0.1549
PHP*Age	0.09972	0.07099	0.1605
Age*SI*Age	-0.05587	0.1516	0.7125
PHP*Unemployed	0.1234	0.1076	0.2521
PHP*PHP	-0.02395	0.03017	0.4276
PHP*SI*Unemployed	0.02207	0.1154	0.8484
Age*SI*Unemployed	0.07259	0.1576	0.6452
Age*SI*Age*SI	-0.02124	0.08134	0.7940
Susceptibility probability			
Intercept	0.2423	0.1903	0.2031
Age	1.7107	0.2147	$<.0001$
Unemployed	0.4928	0.2193	0.0249
SI	0.2117	0.1506	0.1601
Age*SI	-0.2673	0.1901	0.1600
Dispersion $\log(\kappa)$			
Intercept	-0.1059	0.1110	0.3404
Summary statistics			
Max logL		-1864.4	
AIC		3778.8	

Web Appendix D: Sample SAS codes

```
ods output CovMatParmEst=CoV_Mat_marg;
title 'Method 1: Direct Marginal approach';
proc nlmixed data=data_temps1 GCONV=1e-20 fCONV=1e-20
ABSFCONV=1e-20  ABSGCONV=1e-20 cov;

parms  beta0-beta5=0 gamma0-gamma4 = 0 log_k1   =  0;

k1=exp(log_k1);
response= d1mfs ; /*d1s d2s d1mfs d2mfs dfs1 dfs2*/
eta_marg =   beta0
           + beta1*age1
           + beta2*employ_no
           + beta3*oh2
           + beta4*sugar_intake
           + beta5*age1*sugar_intake ;

eta_inf =    gamma0
           + gamma1*age1
           + gamma2*employ_no
           + gamma3*sugar_intake
           + gamma4*age1*sugar_intake ;

p=exp(eta_inf)/(1+exp(eta_inf));
mu_marg = exp(eta_marg);
mu_cond =mu_marg/p;
if response=0 then loglik= log((1-p)+(p)*exp(-(1/k1)*log(1+mu_cond*k1)));
else loglik=log(p)+lgamma(response+(1/k1))-lgamma(1/k1)-lgamma(response+1)
           -(response+(1/k1))*log(1+mu_cond*k1)+response*log(mu_cond*k1);
ll=loglik;
f0=exp(-(1/k1)*log(1+mu_cond*k1));
if response=0 then p_updated=p*f0/(p*f0+1-p); else p_updated=1;
predict p_updated out=p_updated_marg;
exp_cond=log(mu_cond);
predict exp_cond out=predict_cond_mean1;
model response ~ general(ll);
run;

ods output CovMatParmEst=CoV_Mat;

title 'Method 2: Indirect Marginal approach, computing the marginal
                 parameter from the conditional ones';

proc nlmixed data=data_temps1   GCONV=1e-20 fCONV=1e-20  ABSFCONV=1e-20
     ABSGCONV=1e-20 cov;

parms alpha0-alpha5=0 gamma0-gamma4 = 0 log_k1   =  0;

k1=exp(log_k1);
response=  d1mfs; /*d1s d2s d1mfs d2mfs dfs1 dfs2*/
```

```
eta_cond =    alpha0
            + alpha1*age1
            + alpha2*employ_no
            + alpha3*oh2
            + alpha4*sugar_intake
            + alpha5*age1*sugar_intake;

eta_inf =    gamma0
           + gamma1*age1
           + gamma2*employ_no
           + gamma3*sugar_intake
           + gamma4*age1*sugar_intake;

p=exp(eta_inf)/(1+exp(eta_inf));
mu_cond = exp(eta_cond);

/* NB model */

if response=0 then
    loglik= log( (1-p) + (p)*exp(-(1/k1)*log(1+mu_cond*k1))) ;
else loglik= log(p) + lgamma(response+(1/k1)) - lgamma(1/k1) -
             lgamma(response+1) - (response+(1/k1))*log(1+mu_cond*k1)
             + response*log(mu_cond*k1);

mu_marg=p*mu_cond;
ll=loglik;
f0=exp(-(1/k1)*log(1+mu_cond*k1));
if response=0 then p_updated=p*f0/(p*f0+1-p); else p_updated=1;
model response ~ general(ll);
predict p_updated out=p_updated_cond;
predict eta_cond out=predict_cond_mean2;
predict mu_marg out=marginal_mean;
predict p out=mixing_weight;
run;


proc IML;

use marginal_mean;
read all var {age1 employ_no oh2 sugar_intake} into x_temp;

read all var {age1 employ_no sugar_intake} into z_temp;
read all var {pred} into marg_mean;

use mixing_weight;
read all var {pred} into p;

use Cov_Mat;
read all into cov_estimates;

cov_estimates=cov_estimates[, 2:nrow(cov_estimates)];
```

```
n= nrow(x_temp);

X=j(n,1)||x_temp||(x_temp[,1]#x_temp[,4]);
log_marg_mean=log(marg_mean);
beta_hat=inv(X`*X)*X`*log_marg_mean;
print beta_hat;

var_alpha_gamma=cov_estimates[1:11,1:11];
print var_alpha_gamma;

z=j(n,1)||z_temp||(z_temp[,1]#z_temp[,3]);

delta_Cov=j(n,n,0);

do i=1 to n;
  do j=1 to n;
 delta_Cov[i,j]=((X[i,])||((1-p[i,])*Z[i,]))*var_alpha_gamma
                *((X[j,])||((1-p[j,])*Z[j,]))`;
 end;
end;

Var_Cov_beta_hat=(inv(X`*X)*X`)*delta_Cov*(inv(X`*X)*X`)`;

print Var_Cov_beta_hat;

variances_temp=diag(Var_Cov_beta_hat);
variances=variances_temp(|,+|);
Std_err=sqrt(variances);
Z_value=beta_hat/std_err;
P_value=1-probchi(z_value#z_value,1);
Estimate=beta_hat;
Parameter={"Intercept" "age1" "employ_no" "oh2" "sugar_intake"
           "age1*sugar_intake"}`;

print Parameter Estimate Std_Err Z_value p_value;

quit;
```